SUPPLEMENTARY MATERIAL

Lysine carboxylation: Unveiling a spontaneous post-translational modification

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

a. Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, 60607, USA

b. Children's National Medical Center, Center for Genetic Medicine Research (CGMR), Washington, DC 20010, USA

SUPPLEMENTARY FIGURE, TABLES, DATA 2

SUPPLEMENTARY FIGURE 2
FIGURE S1, SEQUENCE CONSERVATION NEAR THE KCX RESIDUE. 2

SUPPLEMENTARY TABLES 3
TABLE S1. PROTEINS WITH KNOWN CARBOXYLATED LYSINE RESIDUES INCLUDED IN OUR DATA SET. 3
TABLE S2. FREQUENCIES ON *KCX-SITES*. 5
TABLE S3. FREQUENCIES ON *LYS-SITES.* 6
TABLE S4. PERFORMANCE OF THE BAYESIAN CLASSIFIER (PRELYSCAR). 9
TABLE S5. PREDICTIONS ON THE PDB40 DATABASE 10

SUPPLEMENTARY LISTS 11
LIST S1. PDB90. 11
LIST S2. PDB40. 11
LIST S3. PDB90HR. 11
LIST S4. PDB40HR. 11
LIST S5. PREDICTED KCX RESIDUES ON THE PDB90 WITH KCX90RR 11
LIST S6. PREDICTED KCX RESIDUES ON THE PDB90 WITH KCX40RR 11
LIST S7. PREDICTED KCX RESIDUES ON THE PDB90 WITH KCX251 11

SUPPLEMENTARY NOTES 12

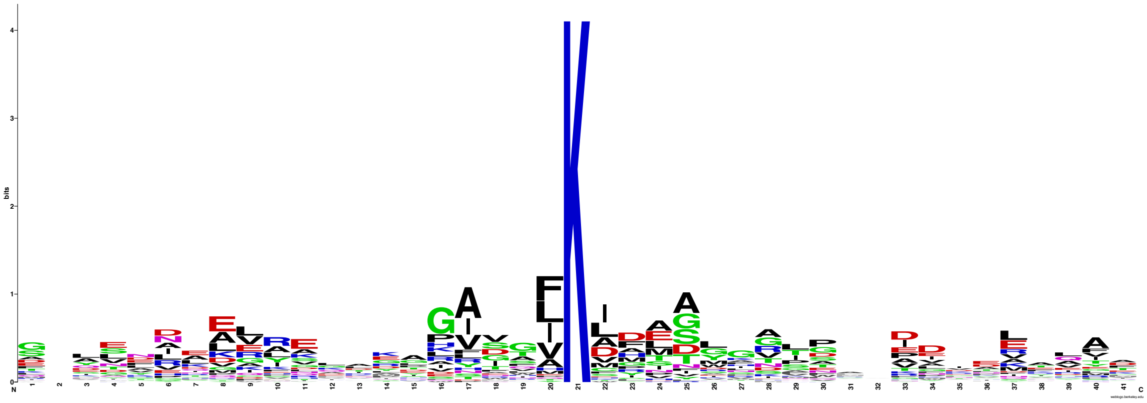
NOTE S1. SEARCHING FOR SEQUENCE MOTIF ON KCX SITES. 12
NOTE S2. PERFORMANCE EVALUATION OF PRELYSCAR. 12
NOTE S3. BIOCHEMICAL FUNCTION OF PROTEINS WITH KNOWN KCX GROUPS. 13
NOTE S4. UNVEILING INCORRECTLY REPORTED KCX RESIDUES. 14
NOTE S5. PREDICTED KCX (PKCX) SITES. 16
PKCX PROTEINS WITH OVERALL SIMILARITY TO KNOWN KCX PROTEINS 16
PKCX PROTEINS WITHOUT OVERALL SIMILARITY TO KNOWN KCX PROTEINS 18

Supplementary Figure, Tables, Data

Supplementary Figure

Figure S1, Sequence conservation near the Kcx residue.

A fragment of 40 amino acids was extracted with the Kcx residue in the center of each protein sequence. The resulting multiple sequence alignment was used to create a sequence logo (Crooks *et al.*, 2004).



Supplementary Tables

Table S1. Proteins with known carboxylated lysine residues included in our data set.

Class of enzyme according to the chemical reaction they catalyze; EC number; PDB ID and the chain in which the carboxylated lysine residue was reported; protein name; the main role of the KCX residue; ion(s); specific metal center motif: parenthesis represent the side chains of the residues interacting with the metal ion. For example, (H,H,D)Zn represents side chains of two His and one Asp interacting with a Zn 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).

Group	EC#	PDB_chain	Protein Name	KCX role	Ion(s)	Metal Center Motif	Specie
Hydrolase	3.5.-.-	3OJG_A	Phosphotriesterase, Lactonase	Metal Ion Center	Zn-Fe	(H,H,N)Zn…KCX…Fe(H,H)	*Geobacillus kaustophilus*
Hydrolase	3.5.1.5	1E9Z_B	Urease	Metal Ion Center	Ni-Ni	(H,H,D)Ni…KCX…Ni(H,H,G)	*Helicobacter pylori*
Hydrolase	3.5.1.5	1EJX_C	Urease	Metal Ion Center	Ni-Ni	(H,H,D)Ni…KCX…Ni(H,H,G)	*Klebsiella aerogenes*
Hydrolase	3.5.2.2	1GKP_C	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Thermus sp*
Hydrolase	3.5.2.2	1GKR_D	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Arthobacter aurescens*
Hydrolase	-	1K1D_C	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Geobacillus stearothermophilus*
Hydrolase	3.5.2.2	1NFG_C	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Ralstonia pickettii*
Hydrolase	3.4.19.-	1ONW_A	Isoaspartyl Dipeptidase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Escherichia coli*
Hydrolase	3.5.2.2	2FTW_A	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Dictyostelium discoideum*
Hydrolase	3.5.2.2	2FVK_C	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Saccharomyces kluyveri*
Unknown	-	2GWN_A	Putative Dihydroorotase	Metal Ion Center	Zn-Zn	(H,Q,D)Zn…KCX…Zn(H,H)	*Porphyromonas gingivalis*
Hydrolase	3.5.4.2	2ICS_A	Adenine Deaminase	Metal Ion Center	Zn-Zn	(H,H,Y)Zn…KCX…Zn(H,H,D)	*Enterococcus faecalis*
Hydrolase	3.1.8.1	2OB3_B	Aryldialkylphosphatase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Brevundimonas diminuta*
Hydrolase	-	2OGJ_F	Dihydroorotase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Agrobacterium tumefaciens*
Hydrolase	-	2QPX_A	Hydrolase, metal-dependent	Metal Ion Center	Zn-Zn	(H,H,Y)Zn…KCX…Zn(H,H,D)	*Lactobacillus casei*
Hydrolase	3.1.8.1	2VC7_B	Aryldialkylphosphatase	Metal Ion Center	Co-Fe	(H,H,D)Fe…KCX…Co(H,H)	*Sulfolobus solfataricus*
Hydrolase	3.5.2.3	2Z26_B	Dihydroorotase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Escherichia coli*
Hydrolase	3.5.2.2	3DC8_B	D-Hydantoinase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Sinorhizobium meliloti*
Hydrolase	-	3DUG_G	Arginine carboxypeptidase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Environmental sample*
Hydrolase	3.5.2.5	3E74_B	Allantoinase	Metal Ion Center	Fe-Fe	(H,H,D)Fe…KCX…Fe(H,H)	*Escherichia coli*
Hydrolase	-	3GTX_A	Organophosphorus hydrolase	Metal Ion Center	Co-Co	(H,H,D)Co…KCX…Co(H,H,Y)	*Deinococcus radiodurans*
Hydrolase	-	3ICJ_A	Metal-Dependent Hydrolase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Pyrococcus furiosus*
Hydrolase	3.5.2.3	3JZE_D	Dihydroorotase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Salmonella enterica*
Hydrolase	3.5.1.5	3LA4_A	Urease	Metal Ion Center	Ni-Ni	(H,H,D)Ni…KCX…Ni(H,H,G)	*Canavalia ensiformis*
Hydrolase	-	3MKV_H	Putative Amidohydrolase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Unidentified*
Hydrolase	-	3MTW_A	L-Lys, L-Arg Carboxypeptidase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Caulobacter vibrioides*
Hydrolase	-	3N2C_P	Putative Amidohydrolase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Unidentified*

Hydrolase	-	3OVG_A	Amidohydrolase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Mycoplasma synoviae*
Hydrolase	3.5.2.3	3PNU_B	Dihydroorotase	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Campylobacter jejuni*
Hydrolase	-	3PNZ_A	Lactonase Lmo2620	Metal Ion Center	Zn-Zn	(H,H,D)Zn…KCX…Zn(H,H)	*Listeria monocytogenes*
Hydrolase	3.5.1.5	3QGK_O	Urease	Metal Ion Center	Fe-Fe	(H,H,D)Fe…KCX…Fe(H,H)	*Helicobacter mustelae*
Hydrolase	3.5.1.5	4UBP_C	Urease	Metal Ion Center	Ni-Ni	(H,H,D)Ni…KCX…Ni(H,H,G)	*Bacillus pasteurii*
Ligase	6.3.2.12	1W78_A	Folc Bifunctional Protein	Metal Ion Center	Mg	Mg(HOH…KCX…HOH)	*Escherichia coli*
Lyase	4.1.1.39	1BWV_C	Rubisco	Metal Ion Center	Mg	(E,D)Mg…KCX	*Galdieria partita*
Transferase	2.1.3.1	1RQB_A	Transcarboxylase 5S Subunit	Metal Ion Center	Co	(H,H,D)Co…KCX	*Propionibacterium freudenreichii*
Lyase	4.1.1.39	1WDD_A	RuBisCo	Metal Ion Center	Mg	(E,D)Mg…KCX	*Oryza sativa*
Isomerase	5.3.2.-	2OEM_B	Rubisco-like protein, enolase	Metal Ion Center	Mg	(E,D)Mg…KCX	*Geobacillus kaustophilus*
Ligase	6.4.1.1	2QF7_A	Pyruvate Carboxylase	Metal Ion Center	Zn	(H,H,D)Zn…KCX	*Rhizobium etli*
Ligase	6.3.2.13	2XJA_A	MurE Ligase	Metal Ion Center	Mg	Mg(HOH…KCX…HOH)	*Mycobacterium tuberculosis*
Ligase	6.4.1.1	3BG3_C	Pyruvate Carboxylase	Metal Ion Center	Mn	(H,H,D)Mn…KCX	*Homo sapiens*
Lyase	4.1.1.39	3KDN_H	RuBisCo	Metal Ion Center	Mg	(E,D)Mg…KCX	*Thermococcus kodakarensis*
Lyase	-	2J6V_A	Uv Damage Endonuclease	Assist uon binding	Mn(3x)	KCX not in metal center	*Thermus thermophilus*
Hydrolase	3.5.2.6	1K38_B	Oxa-2 Class D Beta-Lactamase	Catalytic	-	-	*Salmonella typhimurium*
Hydrolase	3.5.2.6	2X02_B	Oxa-10 Class D Beta-Lactamase	Catalytic	-	-	*Pseudomonas aeruginosa*
Hydrolase	3.5.2.6	3G4P_A	Oxa-24 Class D Beta-Lactamase	Catalytic	-	-	*Acinetobacter baumannii*
Hydrolase	3.5.2.6	3HBR_A	Oxa-48 Class D Beta-Lactamase	Catalytic	-	-	*Klebsiella pneumoniae*
Hydrolase	3.5.2.6	3IF6_C	Oxa-46 Beta-Lactamase	Catalytic	-	-	*Pseudomonas aeruginosa*
Hydrolase	3.5.2.6	3ISG_B	Oxa-1 Class D Beta-Lactamase	Catalytic	-	-	*Escherichia coli*
Ligase	6.3.2.9	2X5O_A	MurD inhibitor	Catalytic?	-	-	*Escherichia coli*
Lyase	-	3NWR_A	Rubisco-Like Protein	Co-catalytic	-	-	*Burkholderia fungorum*
Isomerase	5.1.1.1	1RCQ_A	Alanine Racemase	Hydrogen bonding	-	-	*Pseudomonas aeruginosa*
Isomerase	5.1.1.1	1VFS_A	Alanine Racemase	Hydrogen bonding	-	-	*Streptomyces lavendulae*
Isomerase	5.1.1.1	1XQL_A	Alanine Racemase	Hydrogen bonding	-	-	*Geobacillus stearothermophilus*
Isomerase	5.1.1.1	2ODO_C	Alanine Racemase	Hydrogen bonding	-	-	*Pseudomonas Fluorescens*
Isomerase	5.1.1.1	2RJH_C	Alanine Racemase	Hydrogen bonding	-	-	*Escherichia coli*
Isomerase	5.1.1.1	3S46_A	Alanine Racemase	Hydrogen bonding	-	-	*Streptococcus pneumoniae*
Hydrolase	-	1PU6_B	3-Methyladenine DNA Glycosylase	Hydrogen bonding?	-	-	*Helicobacter pylori*
Ligase	6.3.2.13	1E8C_B	UDP-N-acetylmuramoyl-l-alanine	Not provided	-	-	*Escherichia coli*
Transferase	2.7.11.22	1H01_A	Human Cyclin Dependent Kinase 2	Not provided	-	-	*Homo sapiens*
Ligase	6.3.2.17	2GC5_A	Folylpolyglutamate Synthase	Not provided	-	-	*Lactobacillus casei*
Transferase	2.1.3.9	3KZN_A	N-Acetyl-L-Ornithine Transcarbamylase	Not provided	-	-	*Xanthomonas campestris*
Hydrolase	-	3Q7V_B	Beta-Lactamase Regulatory Protein	Switch	-	-	*Staphylococcus aureus*

Table S2. Frequencies on *KCX-sites*.

Frequency of amino acids, metal ions (ION), and water molecules (WAT) found within 5 Å from the side chain of Lys residues (*KCX-sites*). The amino acids found in the microenvironments are grouped according to their main physicochemical properties, i.e. positively ionizable (POS: Arg, Lys, His), negatively ionizable (NEG: Asp, Glu), aromatic (ARO: Trp, Phe, Tyr), small polar (ST: Ser, Thr), long polar (NQ: Asn, Gln), and hydrophobic (HYD: Ile, Leu, Val, Met) residues.

KCX-sites									
BIN	LEN	NEG	POS	ST	NQ	ARO	HYD	ION	WAT
0	-	0.597	0.016	0.145	0.758	0.113	-	0.371	0.274
1	-	0.323	0.194	0.419	0.226	0.403	0.016	0.113	0.371
2	-	0.081	0.161	0.242	0.016	0.258	0.081	0.516	0.226
3	-	-	0.419	0.065	-	0.21	0.258	-	0.081
4	-	-	0.177	0.097	-	0.016	0.274	-	0.032
5	-	-	0.032	0.032	-	-	0.371	-	0.016
6	-	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-
11	0.032	-	-	-	-	-	-	-	-
12	0.258	-	-	-	-	-	-	-	-
13	0.419	-	-	-	-	-	-	-	-
14	0.194	-	-	-	-	-	-	-	-
15	0.048	-	-	-	-	-	-	-	-
16	0.016	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-

Table S3. Frequencies on *LYS-sites.*

Frequency of amino acids, metal ions (ION), and water molecules (WAT) found within 5 Å from the side chain of (A) all lysine residues (*LYS-sites*), (B) buried lysine residues, and (C) surface residues. The amino acids found in the microenvironments are grouped according to their main physicochemical properties, i.e. positively ionizable (POS: Arg, Lys, His), negatively ionizable (NEG: Asp, Glu), aromatic (ARO: Trp, Phe, Tyr), small polar (ST: Ser, Thr), long polar (NQ: Asn, Gln), and hydrophobic (HYD: Ile, Leu, Val, Met) residues.

(A) All LYS-sites

BIN	LEN	NEG	POS	ST	NQ	ARO	HYD	ION	WAT
0	-	0.218	0.58	0.536	0.58	0.493	0.215	0.991	0.239
1	-	0.421	0.313	0.310	0.306	0.344	0.309	0.007	0.156
2	0.028	0.257	0.086	0.102	0.091	0.121	0.224	0.001	0.130
3	0.070	0.086	0.016	0.036	0.022	0.031	0.146	-	0.137
4	0.125	0.016	0.003	0.013	-	0.01	0.07	-	0.095
5	0.162	-	0.001	0.002	-	0.001	0.036	-	0.243
6	0.144	-	-	0.001	-	-	0.036	-	-
7	0.109	-	-	-	-	-	0.036	-	-
8	0.105	-	-	-	-	-	0.001	-	-
9	0.064	-	-	-	-	-	-	-	-
10	0.048	-	-	-	-	-	-	-	-
11	0.049	-	-	-	-	-	-	-	-
12	0.037	-	-	-	-	-	-	-	-
13	0.025	-	-	-	-	-	-	-	-
14	0.017	-	-	-	-	-	-	-	-
15	0.010	-	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-

(B) Buried LYS-sites

BIN	LEN	NEG	POS	ST	NQ	ARO	HYD	ION	WAT
0	-	0.124	0.380	0.270	0.380	0.328	0.058	0.912	0.219
1	-	0.423	0.299	0.255	0.328	0.328	0.088	0.066	0.161
2	-	0.241	0.234	0.248	0.255	0.241	0.219	0.022	0.124
3	-	0.182	0.073	0.124	0.029	0.058	0.328	-	0.124
4	-	0.029	0.015	0.073	0.007	0.044	0.182	-	0.131
5	0.007	-	-	0.022	-	-	0.080	-	0.080
6	-	-	-	0.007	-	-	0.029	-	0.109
7	0.007	-	-	-	-	-	0.007	-	0.022
8	0.036	-	-	-	-	-	0.007	-	0.022
9	0.073	-	-	-	-	-	-	-	0.007
10	0.131	-	-	-	-	-	-	-	-
11	0.168	-	-	-	-	-	-	-	-
12	0.212	-	-	-	-	-	-	-	-
13	0.190	-	-	-	-	-	-	-	-
14	0.102	-	-	-	-	-	-	-	-
15	0.051	-	-	-	-	-	-	-	-
16	0.015	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-
18	0.007	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-

(C) Surface LYS-sites

BIN	LEN	NEG	POS	ST	NQ	ARO	HYD	ION	WAT
0	-	0.235	0.611	0.584	0.620	0.532	0.243	0.983	0.253
1	0.003	0.424	0.315	0.313	0.295	0.341	0.345	0.017	0.159
2	0.033	0.254	0.065	0.076	0.068	0.099	0.220	-	0.131
3	0.083	0.072	0.008	0.021	0.016	0.025	0.118	-	0.141
4	0.146	0.014	0.001	0.005	0.001	0.003	0.053	-	0.091
5	0.189	0.001	-	-	-	-	0.015	-	0.073
6	0.169	-	-	-	-	-	0.005	-	0.041
7	0.127	-	-	-	-	-	0.000	-	0.037
8	0.116	-	-	-	-	-	0.001	-	0.026
9	0.062	-	-	-	-	-	-	-	0.013
10	0.033	-	-	-	-	-	-	-	0.022
11	0.027	-	-	-	-	-	-	-	0.004
12	0.010	-	-	-	-	-	-	-	0.006
13	0.002	-	-	-	-	-	-	-	0.001
14	0.002	-	-	-	-	-	-	-	0.001
15	0.001	-	-	-	-	-	-	-	0.001
16	-	-	-	-	-	-	-	-	0.001
17	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-

Table S4. Performance of the Bayesian classifier (PreLysCar).

Leave one-out cross validation test using different prior probability values. The **kcx90rr** data set consists of 62 *KCX-sites* and 1,337 *LYS-sites*. The **kcx40rr** data set consists of 43 *KCX-sites* and 954 *LYS-sites.* The **kcx251** consists in the entire KCX protein data set without redundancy removal (251 *KCX-sites* and 4,259 *LYS-sites*). Abbreviations: SEN = Sensitivity; SPF = Specificity; ACU = Accuracy; PPV = Positive Predictive Value; NPV = Negative Predictive Value; MCC = Matthews Correlation Coefficient; TP = True Positives; TN = True Negative; FP = False Positive; FN = False Negative.

	Prior	**SEN**	**SPF**	**ACU**	**PPV**	**NPV**	**MCC**	**TP**	**TN**	**FP**	**FN**
	0.5	0.936	0.969	0.968	0.586	0.997	0.726	58	1296	41	4
kcx90rr	0.05	0.903	0.990	0.986	0.800	0.996	0.843	56	1323	14	6
	0.009	0.871	0.997	0.991	0.931	0.994	0.896	54	1333	4	8

	Prior	**SEN**	**SPF**	**ACU**	**PPV**	**NPV**	**MCC**	**TP**	**TN**	**FP**	**FN**
	0.5	0.884	0.970	0.966	0.567	0.995	0.692	38	926	29	5
kcx40rr	0.05	0.861	0.992	0.986	0.822	0.994	0.834	37	947	8	6
	0.009	0.837	0.996	0.989	0.900	0.993	0.862	36	951	4	7

	Prior	**SEN**	**SPF**	**ACU**	**PPV**	**NPV**	**MCC**	**TP**	**TN**	**FP**	**FN**
	0.5	0.96	0.973	0.972	0.677	0.998	0.793	241	4144	115	10
kcx251	0.05	0.92	0.991	0.987	0.862	0.995	0.884	231	4222	37	20
	0.009	0.88	0.999	0.992	0.982	0.993	0.926	221	4255	4	30

Table S5. Predictions on the PDB40 database

Predictions on the PDB database. PreLysCar was executed on a subset obtained from the PDB database with a 40% redundancy reduction (PDB40, see Methods). The predicted KCXs (pKCX) on the high-resolution fraction (<1.5 Å) were assumed incorrect and the False Discovery Rate (FPR) calculated. The FPR was further used to estimate the expected error (*e*-error) on the protein structures with a resolution larger than 1.5 Å, and the corresponding expected number of correct predictions (*e*-correct).

B. Predictions on the PDB database

	PDB40 <1.5 Å		PDB40 >1.5 Å		
	pKCX	FPR	pKCX	*e*-error	*e*-correct
kcx40rr	13 / 575	3.4%	411 / 8508	289	122
kcx90rr	7 / 575	1.8%	335 / 8508	153	182
kcx251	10 / 575	2.6%	376 / 8508	221	155

Supplementary Lists

List S1. PDB90.

List of protein chains, with a resolution larger or equal than 1.5 Å, more than 199 amino acid length, and after 90% redundancy reduction (n = 14,262).

List S2. PDB40.

List of protein chains, with a resolution larger or equal than 1.5 Å, more than 199 amino acid length, and after 40% redundancy reduction (n = 8,508).

List S3. PDB90hr.

List of protein chains with a resolution less than 1.5 Å, more than 199 amino acid length, and after 90% redundancy reduction (n = 575).

List S4. PDB40hr.

List of protein chains with a resolution less than 1.5 Å, more than 199 amino acid length, and after 40% redundancy reduction (n = 381).

List S5. Predicted KCX residues on the PDB90 with kcx90rr

List of protein chains with a predicted carboxylated lysine residue (pKCX) on the 90RR data set (number of pKCX proteins = 543).

List S6. Predicted KCX residues on the PDB90 with kcx40rr

List of protein chains with a predicted carboxylated lysine residue (pKCX) on the 90RR data set (number of pKCX proteins = 659).

List S7. Predicted KCX residues on the PDB90 with kcx251

List of protein chains with a predicted carboxylated lysine residue (pKCX) on the 90RR data set (number of pKCX proteins = 601).

Supplementary Notes

Note S1. Searching for sequence motif on KCX sites.

We investigated whether a sequence motif could be identified at the KCX site. Amino acid subsequences around carboxylated lysine residues were first extracted (20 amino acids in each direction from the Kcx residue). We then followed two different approaches. First, we used MEME (Bailey *et al.*, 2009) to search for motifs using default parameters. No significant motifs were found. Second, we measured the sequence conservation at every position by calculating the sequence entropy and created a sequence logo using WebLogo (Crooks *et al.*, 2004). A lack of sequence conservation in the amino acids surrounding the KCX site was also observed (**Figure S1**).

Note S2. Performance evaluation of PreLysCar.

The performance of the Bayesian model was evaluated with the technique of leave-one-out cross-validation. Briefly, a testing vector is held out from the *KCX* and *LYS-site* data sets. The remaining vectors are used for training, from which *KCX* and *LYS-sites* frequencies are obtained. Posterior probabilities for the vector held out are next calculated, both $p(C_{KCX}|F_1,...,F_n)$, i.e., probability of lysine residue of being carboxylated given the composition of its microenvironment, and $p(C_{KCX}|F_1,...,F_n)$, i.e., probability of lysine residue of being uncarboxylated given the composition of its microenvironment. The predictor assigns the class “KCX” or “LYS” according to the largest posterior probability value. This operation is repeated for every vector in the entire data set. As a result, the four different possible outcomes of a binary prediction method are obtained, i.e., total number of “True Positives” (TP, KCX correctly predicted as KCX), “True Negatives” (TN, LYS correctly predicted as LYS), “False Positives” (FP, LYS incorrectly predicted as KCX) and “False Negatives” (FN, KCX incorrectly predicted as LYS). Finally, common statistical measures of performance are calculated, i.e., sensitivity (SEN), specificity (SPE), accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV), Matthews Correlation Coefficient (MCC), and False Positive Rate (fall-out, FPR).

$$SEN = \frac{TP}{TP + FN}$$

$$SPE = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

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

$$FPR = \frac{FP}{FP + TN}$$

Several leave-one-out-cross validation tests were carried out using different prior probabilities. For example, for the 90RR data set (**Table S4**):

- *Prior* = 0.5, an unrealistic assumption of both Kcx and Lys having the same probability of occurrence. Under this prior probability value, large number of correctly predicted *KCX-sites* should be expected (high sensitivity), but also a large number of *LYS-sites* incorrectly classified as *KCX-sites* (low specificity). However, this was not the outcome and 41 false positives out of 1,337 were obtained. Both sensitivity and specificity were above the 90%, with an accuracy of 97%.

- *Prior* = 0.05, probability of finding a KCX residue from the total number of lysine residues in the 90RR data set (62 KCX sites out of 1,399 *KCX+LYS sites*). Such small prior probability value should improve the specificity by reducing the number of false positives, but worsen the sensitivity with less *KCX-sites* correctly predicted. As expected, the number of false positives was reduced, improving from 41 to 14 out of 1,337, but the number of *KCX-sites* correctly predicted was slightly affected (58 to 56). As a result, both sensitivity and specificity of 0.90 and 0.99, respectively, which translates in a better accuracy (0.99).

- *Prior* = 0.009, a generous approximation to the probability of finding a KCX residue in the entire lysine universe. As expected, the very low probability value causes a decrease in the number of correctly predicted *KCX-sites*, but not in a significant number (54 out of 62 *KCX-sites*), which maintains a high sensitivity (0.87). The number of correctly predicted *LYS-sites* was very high (1,333 out of 1,337), which results in a specificity of 0.997.

Note S3. Biochemical function of proteins with known KCX groups.

We studied the enzymatic functions of proteins with reported Kcx residues. We used the EC number available in the PDB when available, i.e. the classification scheme according to the

Enzyme Commission (Tipton, 1994) to describe the biochemical function of enzymes solved with a carboxylated lysine residue.

About 80% of proteins with a known Kcx residue in our data set belong to eubacteria, 6% to eukaryotes, and 3% to archaea. The largest group of enzymes with known carboxylated lysine residues are hydrolases (EC 3). Among them, class D β-lactamases are involved in the resistance to β-lactam antibiotics (Poirel *et al.*, 2010). Ureases are another class of hydrolases with *KCX-sites* found in many bacterial pathogens, and are important for their clinical detection (Mobley *et al.*, 1995). Hydrolases aryldialkylphosphatases (LeJeune *et al.*, 1998) are of great interest due to their potential use in detoxification of chemical waste and warfare agents (LeJeune *et al.*, 1998).

Among the ligases (EC 6), the UDP-N-acetylmuramoyl-L-alanine-D-glutamate ligase is an enzyme involved in the synthesis of the bacterial peptidoglycan (Barreteau *et al.*, 2008). Pyruvate carboxylase is a multifunctional enzyme catalyzing the biotin-dependent production of oxaloacetate with important roles in gluconeogenesis, lipogenesis, insulin secretion, and other cellular processes (St Maurice *et al.*, 2007).

Among transferases (EC:2), the 5S metalloenzyme chain of the transcarboxylase multienzyme (Hall *et al.*, 2004) requires a carboxylated lysine to coordinate a cobalt ion in the active site. Another transferase with a Kcx residue is the N-acetyl-L-ornithine transcarbamoylase (AOTCase). This is an essential enzyme for arginine biosynthesis in several eubacteria (Shi *et al.*, 2006).

The most remarkable lyase (EC:4) known to require lysine carboxylation is the ribulose bisphosphate carboxylase/oxygenase (rubisco), enzyme responsible for creating organic carbon from the inorganic carbon dioxide of the air (Berg *et al.*, 2002).

Alanine racemase is an isomerase (EC:5) that requires a Kcx residue to stabilize the active site. This enzyme catalyzes a step involved in bacterial cell wall biosynthesis by the interconversion of alanine enantiomers (Strych *et al.*, 2000). It is also an attractive target for antimicrobial drug development because of the lack of any homologs in eukaryotes (Silverman, 1988).

Note S4. Unveiling incorrectly reported KCX residues.

We initially began testing PreLysCar on a *KCX-site* data set that contained three more proteins with lysine residues reported as carboxylated. These are the TdcF of *Escherichia coli*, alpha-l-fucosidase of *Thermotoga maritima*, and class C beta-lactamase of *Enterobacter cloacae*. PreLysCar classified them as LYS. After a detailed analysis, we believe that these proteins should be considered uncarboxylated.

TdcF from *Escherichia coli* (PDB 2UYN).

TdcF is a member of the highly conserved family YjgF/YER057c/UK114, widespread in nature. The biological function is not known (Burman *et al.*, 2007). Kcx58 of TdcF (PDB 2UYN) is likely to be a non-carboxylated lysine residue. First, the authors suggested that the KCX modification could be an artifact with no biological importance (Burman *et al.*, 2007). Second, the ambiguity of the electron density map at the tip of Lys58 compelled the authors to model the carboxylated lysine in two different positions. Third, known *KCX-sites* 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 resolved incorrectly as carboxylated. As a consequence, it was eliminated from the *KCX-site* data set.

Putative α-l-fucosidase of *Thermotoga maritima* (PDB:1HL9)

α-l-fucosidases catalyze the removal of non-reducing terminal l-fucose residues (Sulzenbacher *et al.*, 2004). Lys338 was resolved with a carboxyl group, although several reasons suggest that this residue is likely non-carboxylated. First, Lys338 showed extra density in chain A, but not in chain B. Second, the remaining atomic coordinates solved in the same publication in different conditions for the same protein, showed low resolution for the side chain of Lys338 (Sulzenbacher *et al.*, 2004). Third, although known *KCX-sites* are found mostly buried, Lys338 is at the surface and quite accessible to solvent. Four, the structure of the same protein was solved years later and Lys338 was not modeled carboxylated (Wu *et al.*, 2009). Finally, fucosidase is a lysosomal enzyme, which operates at pH 5. Given the lability of the carboxyl group under acidic environments, it is unlikely that Lys338 could become carboxylated. Since lysine residues can also be modified by other PTMs, which is difficult to distinguish based on the electron density alone, other PTM cannot be ruled out. Therefore, we believe that Lys338 is likely uncarboxylated. As consequence, it was eliminated from the *KCX-site* data set.

Class C β-lactamase of *Enterobacter cloacae* (PDB 2P9V)

This serine-dependent enzyme is involved in the hydrolysis of the β-lactam ring of the β-lactam antibiotic (Bush *et al.*, 1995). Lys315, resolved as carboxylated but predicted as uncarboxylated, adopted an unusual conformation. A carbamate cross-linking between Lys315 and Ser64 occurred as consequence of the action of an inhibitor (Wyrembak *et al.*, 2007). However, carboxylation does not occur in functional proteins of the class C of β-lactamases, as it occurs only in class-D (Poirel *et al.*, 2010). There are 66 solved structures of class C β-lactamase available at the PDB database sharing more than 98% sequence identity with 2P9V and none of them were resolved with a Kcx residue. Carboxylation occurred in this class C β-lactamase as a consequence of an inhibitor of Lys315 as reported (Sulzenbacher *et al.*, 2004), but does not occur in functional states of the class C β-lactamase. Therefore, PreLysCar correctly classified Lys315 as uncarboxylated. Consequently, the protein was removed from the *KCX-site* data set.

Note S5. Predicted KCX (pKCX) sites.

In this section, we extend the description for some examples of proteins predicted to have a carboxylated lysine residue.

pKCX proteins with overall similarity to known KCX proteins

Phosphotriesterase from *Pseudomonas diminuta* (PDB 1PSC):

Lys169 was predicted as carboxylated. Phosphotriesterase detoxifies paraoxon and parathion, pesticides widely used, in addition to various mammalian acetylcholinesterase inhibitors (Benning *et al.*, 1995). Lys169 was reported as carboxylated and bridging two atoms of cadmium at the active site of this holoenzyme (Benning *et al.*, 1995). However, it does not appear as carboxylated in the atomic coordinates.

Carbapenemase OXA-24 from *Acinetobacter baumannii* (PDB 2JC7):

Lys84 was predicted as carboxylated. Carbapenemase OXA-24, a class D β-lactamase, was isolated from a multi-resistant epidemic clinical strain of *Acinetobacter baumannii* (Santillana *et al.*, 2007). This family of anti-β-lactams is well known to have a carboxylated lysine as part of the active site. The Fo-Fc electron density map shows a clear cloud of electrons at the tip of Lys84. A carboxyl group can be built with confidence. We remodeled Lys84 to Kcx84, obtaining a better fit with the carboxyl group (**Fig. 3**).

Amidohydrolase Sgx9260c from environmental sample (PDB 3FEQ)

Lys188 was predicted as carboxylated. The structure of Sgx9260c was solved along with Sgx9260b (PDB 3MKV), a homologous protein from the same super-family. Both proteins share a 98% sequence identity and their active sites are almost identical, which is typical of type I binuclear metal centers in the amidohydrolase superfamily (Seibert & Raushel, 2005). The active site consists in a zinc-binding site composed of six conserved residues, including a carboxylated lysine. While Sgx9260b (PDB 3MKV) was solved with the carboxylated lysine at position 191, Sgx9260c (PDB 3FEQ) lack the post-translational modification at the equivalent position Lys188, despite our prediction. This is because the structure was solved in the presence of an inhibitor and consequently the metal-binding site of Sgx9260c was only partially occupied by Zn^{2+} and partially filled with water (Xiang *et al.*, 2010).

Uncharacterized metal-dependent hydrolase from *Pyrococcus horikoshii* (PDB 3IGH)

Lys278 was predicted as carboxylated. This protein shows a 70% sequence identity with another metal-dependent hydrolase from *Pyrococcus furiosus* (PDB 3ICJ), where Lys294 was found to be carboxylated and bridging two Zn ions. Both pKCX and KCX sites show a similar microenvironment with one Glu and 5 His residues characteristic of KCX metal ion centers. The predicted pKCX site contains a sulfate ion, which could prevent carboxylation.

Human dihydropyrimidinase (PDB 2VR2)

Lys159 from the human dihydropyrimidinase was predicted as carboxylated. The electron density map shows high density between the tip of the Lys159 and a zinc ion. This protein shares 60% sequence identity with the dihydropyrimidinase from *Dictyostelium discoideum* (PDB 2FTW), which exhibits a Kcx residue at the active site (Kcx158). Dihydropyrimidinases catalyses the reversible hydrolytic ring-opening of cyclic diamides. The active site of dyhydropyrimidinases is characterized for a carboxylated lysine residue involved in bridging a binuclear zinc center.

Alanine racemase from *Enterococcus faecalis* (PDB 3E5P)

Lys132 was predicted as carboxylated. This racemase shares 50% sequence identity with the alanine racemase from *Streptococcus pneumoniae* (PDB 3S46), where Lys129 was resolved as carboxylated and hydrogen bonded to the neighboring arginine residue (Im & Roux, 2002). There are strong envidences for the presence of such carboxylated residue in alanine racemases (LeMagueres *et al.*, 2003; Morollo *et al.*, 1999).

Allantoinase from *Bacillus halodurans* (PDB 3HM7)

Lys150 was predicted as carboxylated. This protein shares 40% sequence identity with the allantoinase of *Escherichia coli* (Kcx146, PDB 3E74 (Kim *et al.*, 2009)). Allantoinases are involved in the final step of the biogenesis and degradation of ureides by catalyzing the conversion of (S)-allantoin into allantoate (Kim *et al.*, 2009). In the allantoinase of *E. coli*, Lys146 is carboxylated and bridging two iron metal ions, with one of the ions coordinated in a DHH...KCX...HH motif. In the case of *Bacillus halodurans* (PDB 3HM7), a zinc atom is present in structural region similar to *E. coli* protein, with a similar motif DHH coordinating the Zn ion. The predicted KCX residue is in the proximity, in addition to other His residues.

Rubisco-like enzymes

PreLysCar identified several protein enzymes with different degrees of similarity to Rubisco proteins that are known to have a carboxylated lysine residue. For example, Lys189 of the type III Rubisco from the hyperthermophilic archaeon *Thermococcus kodakarensis* (Tk-Rubisco,PDB 1GEH) was predicted as carboxylated. Although this residue was resolved as carboxylated (Kitano *et al.*, 2001), it was modeled without the carboxyl group. The same protein was years later solved with the carboxyl group on it (PDB 3A12) (Nishitani *et al.*, 2010). Lys186 of the Rubisco from *Pyrococcus horikoshii* (PDB 2CWX) was also predicted as carboxylated. This Rubisco shows a 40% sequence identity with Tk-Rubisco. Both microenvironment are very similar.

Carboxypeptidase (PDB-ID:2QS8)

Lys194 of this carboxypeptidase from the amidohydrolase superfamily (unknown source) was predicted to be carboxylated. The protein shares a 36% sequence identity with the ZN-dependent arginine carboxypeptidase (PDB 3DUG), which was found to have a carboxylated lysine residue in the equivalent position (Kcx182). Both structures were published simultaneously (Xiang *et al.*, 2010). Despite the low sequence identity, both proteins share the same fold. A binuclear metal center coordinated by Asp and five His residues was expected, with Zn ions bridged by a carboxylated lysine (Xiang *et al.*, 2010). However, they were not able to obtain the metal ions in that position. The authors emphasized the difficulties in capturing this post-translational modification through purification and crystallization. Our prediction confirms their expectations.

Imidazolonepropionase enzymes

KCX residues were predicted in a group of imidazolonepropionase enzymes from three different organisms: *Agrobacterium tumefaciens* (Lys155, PDB 2GOK), *Bacillus subtilis* (Lys149, PDB 2BB0), and from an unknown environmental sample (Lys 141,PDB 2OOF). These three hydrolases share around 40% sequence identity among themselves. The predicted KCX residues are located in a similar position in the three proteins, i.e., buried in a cavity, which is also located in a central region of the protein. Each of these enzymes shared an overall 20 to 30% sequence identity with two other different hydrolases that are known to have carboxylated lysine residues, the D-hydantoinase from *Burkholderia pickettii* (PDB 1NFG) and the amidohydrolase (PDB 3MKV) from an unknown environmental sample (Xiang *et al.*, 2010). Despite the remote homology, all these proteins show similar topology of the active site and overall protein fold.

pKCX proteins without overall similarity to known KCX proteins

Putative oxidoreductase from *Erwinia carotovora atroseptica* (PDB 2P2S)

Lys96 was predicted as carboxylated. This lysine residue is buried within the protein, in the center of a pocket. The microenvironment and the disposition of the residues are similar to other KCX sites. Four His and one Asp residues are within a distance that could allow the formation of a metal ion center. Lys96 was solved forming hydrogen bonds with 3 water molecules and Asp178, which was modeled in two possible conformations, despite the high-resolution of the structure (1.25 Å).

This putative oxidoreductase shows high sequence similarity with a large number of annotated oxidoreductases. Structurally, two main protein domains are identified by PFAM (Punta *et al.*, 2012) that belong as well to the oxidoreductase family. Interestingly, we detected remote sequence similarity in a fragment of 75 residues between this oxidoreductase and the rubisco-like protein from *Burkholderia fungorum* (PDB 3NWR). The fragment covers both the predicted pKcx96 of the oxidoreductase and the carboxylated lysine (Kcx195) of the rubisco-like protein. Although the topology of both proteins is different, the fragments share local structural similarity, consisting of two alpha helices and a region part of the pocket where both Kcx195 and pKcx96 are located.

Class II fructose-biphosphate aldolase from *Helicobacter pylori* (PDB 3C4U)

Lys 251 was predicted as pKCX. The composition and structural organization of the microenvironment resemble the *KCX-sites* described. The lysine predicted as carboxylated is buried in the center of a cavity with one Glu and four His residues within a good distance for metal binding. The protein was solved forming hydrogen bonding with Gln, Asn, and Glu. A sodium ion was found on the tip of the lysine residue, although a negative peak shows in the electron density map for Na. A Zn ion was also found in the proximities of the pKCX coordinated by two His residues.

We detected a 27% sequence identity in a fragment of 80 residues between this aldolase and two different KCX proteins, the OXA 10 class D beta-lactamase from *Pseudomonas aeruginosa* (PDB 2X02) and the transcarboxylase from *Propionibacterium shermanii* (PDB 1RQB). Although the overall topology between the proteins is different, the fragments share local structural similarity, which forms several alpha helices and a region part of the pocket where the KCX and pKCX are located.

Tagatose 6-phosphate kinase from *Escherichia coli* (PDB 2FIQ)

Lys279 was predicted as carboxylated. The microenvironment of this lysine shows the KCX signature, with the pKCX residue situated in the center of the TIM-barrel domain of the protein. Three His and a negatively charged residue are within 5 Å from the amino terminal group of the pKCX. The protein structure was solved at 2.23 Å resolution and pH 6.5. At the moment of writing this article, there is no structural and functional characterization of the mechanism for tagatose 6-phosphate kinases.

Mannose 6-Phosphate Isomerase from *Bacillus subtilis* (PDB 1QWR)

Lys96 was predicted as KCX. This protein does not share sequence similarity with any protein with a known carboxylated lysine residue. This monomeric enzyme catalyzes the interconversion of fructose 6-phosphate and mannose-6-phosphate (Yeom *et al.*, 2009). It uses zinc as cofactor ligand, which appears to play a role in substrate binding and in maintaining the architecture of the active site (Sagurthi *et al.*, 2009). The residue Lys96 lies near a zinc ion in the protein structure. The structure was solved at 1.8 Å and pH of 5.5, which might have prevented carboxylation. Despite the high-resolution, the electron density for the side chain of this amino acid is poor. A known KCX protein was solved with one atom of Zn, the multifunctional pyruvate carboxylase from *Rhizobium etli* (PDB 2QF7), where the Zn atom was found coordinated by 2 His and 1 Asp residues, instead of Glu observed for the mannose 6-phosphate isomerase.

Mandelate racemase from Roseobacter denitrificans (PDB 3TCS) and Roseovarius nubinhibens (PDB 3U4F)

Lys145 is predicted in both proteins to be carboxylated. These two racemases from the muconate lactonizing protein family share an 85% sequence identity between themselves. No sequence similarity was detected to any protein with a known carboxylated lysine residue. However, the microenvironment of both lysine is similar to carboxylated lysine residues described. In both pKCX sites, an atom of Mg is present, although in different conditions. In one of the structures is bind to a D-alanine, while in the other is bind to a nucleotide (guanidine). In addition, the structure of *Roseovarius nubinhibens* (PDB 3U4F) was solved using selenomethionine, which may potentially prevent carboxylation on Lys145.

Glycerol dehydrogenase (GlyDH) from *Bacillus stearothermophilus* (PDB 1JPU) and *Sinorhizobium meliloti* (PDB 3UHJ)

Lys277 and Lys298 were predicted as carboxylated in the glycerol dehydrogenase (GlyDH) from *Bacillus stearothermophilus* (PDB 1JPU) and *Sinorhizobium meliloti* (PDB 3UHJ), respectively. Both proteins share a 40% sequence identity, in addition to an overall similar structural topology. These proteins do not share sequence similarity with any protein with a known carboxylated lysine residue. Both lysine residues occupy equivalent positions within the structure. The composition of the microenvironment is similar as well, with at least five His and a Glu residue within a distance that would allow the formation of a ion metal center.

The GlyDH from *B. stearothermophilus* was originally indicated as part of the iron-containing alcohol dehydrogenase family, but curiously, based on the strict dependency on zinc for the protein activity, the entire group was renamed to "family III metal-dependent polyol dehydrogenases". GlyDH from *B. stearothermophilus* (PDB 1JPU) was solved at 1.8 Å resolution (pH not available) with two zinc ions solved in the same area where the pKCX is found. The GlyDH and *Sinorhizobium meliloti* (PDB 3UHJ) was solved at 2.34 Å, pH of 5.5, and only one of the two Zn ions resolved.

YML079w from *Saccharomyces cerevisiae* (PDB 1XE7)

Lys147 was predicted as carboxylated. It is found buried in a central location of the protein structure. No sequence similarity with any protein with a known carboxylated lysine residue was identified. At least four His and one Glu residues are in positions that resemble a *KCX-site*. This protein, of unknown function, was solved at 1.75 Å 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.

ArnB (PmrH) aminotransferase from Salmonella typhimurium (PDB 1MDO)

Lys188 was predicted as carboxylated. This residue is buried in the active site at the center of the protein. No sequence similarity was detected with any protein with a known carboxylated lysine residue. The composition of the microenvironment is similar to the KCX-sites described. The protein was solved at 1.7 Å and pH 5.5. An electron density feature was found to extend from the ligand, which is in contact with Lys188 and that could not be explained (Noland *et al.*, 2002). In addition, Lys188 was proposed to act as a general base to abstract the α-proton, but also involved as a general acid or base in several steps of the enzymatic mechanism.

Supplementary References

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. (2009). *Nucleic Acids Res* **37**, W202-208.

Barreteau, H., Kovac, A., Boniface, A., Sova, M., Gobec, S. & Blanot, D. (2008). *FEMS Microbiol Rev* **32**, 168-207.

Benning, M. M., Kuo, J. M., Raushel, F. M. & Holden, H. M. (1995). *Biochemistry* **34**, 7973-7978.

Berg, J. M., Tymoczko, J. L. & Stryer, L. (2002). *Biochemistry. 5th edition*. New York: W H Freeman.

Burman, J. D., Stevenson, C. E., Sawers, R. G. & Lawson, D. M. (2007). *BMC Struct Biol* **7**, 30.

Bush, K., Jacoby, G. A. & Medeiros, A. A. (1995). *Antimicrob Agents Chemother* **39**, 1211-1233.

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). *Genome Res* **14**, 1188-1190.

Hall, P. R., Zheng, R., Antony, L., Pusztai-Carey, M., Carey, P. R. & Yee, V. C. (2004). *EMBO J* **23**, 3621-3631.

Im, W. & Roux, B. Æ. (2002). *Journal of Molecular Biology* **319**, 1177-1197

Kim, K., Kim, M. I., Chung, J., Ahn, J. H. & Rhee, S. (2009). *J Mol Biol* **387**, 1067-1074.

Kitano, K., Maeda, N., Fukui, T., Atomi, H., Imanaka, T. & Miki, K. (2001). *Structure* **9**, 473-481.

LeJeune, K. E., Wild, J. R. & Russell, A. J. (1998). *Nature* **395**, 27-28.

LeMagueres, P., Im, H., Dvorak, A., Strych, U., Benedik, M. & Krause, K. L. (2003). *Biochemistry* **42**, 14752-14761.

Mobley, H. L., Island, M. D. & Hausinger, R. P. (1995). *Microbiol Rev* **59**, 451-480.

Morollo, A. A., Petsko, G. A. & Ringe, D. (1999). *Biochemistry* **38**, 3293-3301.

Nishitani, Y., Yoshida, S., Fujihashi, M., Kitagawa, K., Doi, T., Atomi, H., Imanaka, T. & Miki, K. (2010). *J Biol Chem* **285**, 39339-39347.

Noland, B. W., Newman, J. M., Hendle, J., Badger, J., Christopher, J. A., Tresser, J., Buchanan, M. D., Wright, T. A., Rutter, M. E., Sanderson, W. E., Muller-Dieckmann, H. J., Gajiwala, K. S. & Buchanan, S. G. (2002). *Structure* **10**, 1569-1580.

Poirel, L., Naas, T. & Nordmann, P. (2010). *Antimicrobial Agents and Chemotherapy* **54**, 24-38.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A. & Finn, R. D. (2012). *Nucleic Acids Res* **40**, D290-301.

Sagurthi, S. R., Gowda, G., Savithri, H. S. & Murthy, M. R. (2009). *Acta Crystallogr D Biol Crystallogr* **65**, 724-732.

Santillana, E., Beceiro, A., Bou, G. & Romero, A. (2007). *Proceedings of the National Academy of Sciences of the United States of America* **104**, 5354-5359.

Seibert, C. M. & Raushel, F. M. (2005). *Biochemistry* **44**, 6383-6391.

Shi, D., Yu, X., Roth, L., Morizono, H., Tuchman, M. & Allewell, N. M. (2006). *Proteins* **64**, 532-542.

Silverman, R. B. (1988). *Journal of enzyme inhibition* **2**, 73-90.

St Maurice, M., Reinhardt, L., Surinya, K. H., Attwood, P. V., Wallace, J. C., Cleland, W. W. & Rayment, I. (2007). *Science* **317**, 1076-1079.

Strych, U., Huang, H. C., Krause, K. L. & Benedik, M. J. (2000). *Curr Microbiol* **41**, 290-294.

Sulzenbacher, G., Bignon, C., Nishimura, T., Tarling, C. A., Withers, S. G., Henrissat, B. & Bourne, Y. (2004). *J Biol Chem* **279**, 13119-13128.

Tipton, K. F. (1994). *European Journal of Biochemistry* **223**, 1-5.

Wu, H. J., Ho, C. W., Ko, T. P., Popat, S. D., Lin, C. H. & Wang, A. H. (2009). *Angew Chem Int Ed Engl* **49**, 337-340.

Wyrembak, P. N., Babaoglu, K., Pelto, R. B., Shoichet, B. K. & Pratt, R. F. (2007). *Journal of the American Chemical Society* **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.

Yeom, S. J., Ji, J. H., Kim, N. H., Park, C. S. & Oh, D. K. (2009). *Appl Environ Microbiol* **75**, 4705-4710.