The data provide information in support of the research article, “The cleavage specificity of the aspartic protease of cocoa beans involved in the generation of the cocoa-specific aroma precursors” (Janek et al., 2016) [1]. Three different protein substrates were partially digested with the aspartic protease isolated from cocoa beans and commercial pepsin, respectively. The obtained peptide fragments were analyzed by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry (MALDI-TOF/TOF-MS/MS) and identified using the MASCOT server. The N- and C-terminal ends of the peptide fragments were used to identify the corresponding in-vitro cleavage sites by comparison with the amino acid sequences of the substrate proteins. The same procedure was applied to identify the cleavage sites used by the cocoa aspartic protease during cocoa fermentation starting from the published amino acid sequences of oligopeptides isolated from fermented cocoa beans.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Type of data Tables
How data was acquired Peptide mixtures obtained by cleavage of different substrate proteins with purified cocoa aspartic protease or pepsin were analyzed by liquid chromatography-MALDI-TOF/TOF-MS/MS using a 4700 proteomics Analyzer (Applied Biosystems, Framingham, MS) of-line coupled with a Ultimate HPLC system and Probot fractionation devise (both Dionex/Thermo, Idstein, Germany). Amino acid sequences of oligopeptides isolated from fermented cocoa beans were taken from the literature.

Data format Analyzed
Experimental factors Samples were prepared by partial digestion of different substrate proteins with purified cocoa aspartic protease or pepsin. Prior to LC-MALDI-MS/MS analyses, the peptide mixtures were modified by reduction and alkylation of cysteine residues with dithiotreitol and iodoacetamide.

Experimental features Generation of oligopeptide mixtures by digestion of substrate proteins with purified cocoa aspartic protease or pepsin, fractionation and sequencing of the peptides by LC-MALDI-TOF/TOF-MS/MS and subsequent identification of the cleavage sites. Data were compared with the cleavage sites predicted from the sequences of oligopeptides isolated from fermented cocoa beans and analyzed by liquid chromatography-tandem mass spectrometry. The abundance of the different amino acid residues in the P4-P4' positions around the cleavage sites were analyzed to get an insight into the particular cleavage specificity of the cocoa aspartic protease.

Data source location Berlin, Germany, and Jena, Germany
Data accessibility Data are within this article.

Value of the data

1. These data characterize the cleavage sites of the cocoa aspartic protease.
2. Characterization of the cleavage specificity of an endoprotease requires the comparative analysis of the amino acid sequences around many of its cleavage sites.
3. We provide a strategy enabling the discrimination between specific and unspecific cleavage sites of an endoprotease.
4. Our data demonstrate the limitation of the identification of protease cleavage sites by LC-MALDI-TOF/TOF-MS/MS versus ESI-MS/MS.
5. These data will contribute to our knowledge concerning the formation of the cocoa-specific aroma precursors.

1. Data

Three tables are presented. Table 1 contains the cleavage sites in different substrate proteins used by the cocoa aspartic protease and pepsin, respectively, identified by in-vitro proteolysis. Table 2 shows the putative cleavage sites of the cocoa aspartic protease used during commercial cocoa fermentation. Table 3 shows the abundance of the different amino acids in the P4 to P4' positions around the cleavage sites used by the cocoa aspartic protease during in-vitro proteolysis and cocoa fermentation, respectively.
| Substrate                        | Cleavage sites specific for the cocoa protease | Common cleavage sites of cocoa protease and pepsin | Cleavage sites specific for pepsin |
|---------------------------------|-----------------------------------------------|---------------------------------------------------|----------------------------------|
| **Myoglobin**                   |                                               |                                                   |                                  |
| (SwissProt no. P68082)          |                                               |                                                   |                                  |
| Myoglobin                       | EWQQVNLNV 7–14                               | DGEWQQVQL 5–12                                    | GEWQQVQL 6–13                    |
|                                 | WQQVNLNVV 8–15                               | QQVNLNVWG 9–16                                    | VLNWVGKVE 11–18                  |
|                                 | FDKFIKHLK 44–51                               | LNWVKGKE 12–19                                    | NVWGKVEA 13–20                   |
|                                 | LKTEIEMK 50–57                                | HGQEVJLR 25–32                                    | GVKEAADIA 16–23                  |
|                                 | EDLIIKHT 60–67                                | QGEVJLRFL 26–33                                   | KVEADIAG 17–24                   |
|                                 | AVIIHLHS 111–118                              | QEVJLRFL 27–34                                    | VEADIAGH 18–25                   |
|                                 | IHVLSH 113–120                                | LFLLFTGHL 30–37                                   | AGHGOVEQ 23–30                   |
|                                 | HLHLSH 114–121                                | TVVLJIALG 67–74                                   | GHQVEQL 24–31                    |
|                                 | VHLHSH 115–122                                | PIKYLEFI 101–108                                  | EVLJLFT 28–35                    |
|                                 | HPGDFGAD 120–127                              | KYLEFISD 103–110                                  | PETLEKFD 38–45                   |
|                                 | FRNDIAAK 139–146                              | YLEFISDA 104–111                                  | HLKTEAEM 49–56                   |
|                                 | AKYKIELGF 145–152                             | FISDIAIHH 107–114                                 | KTEAEMKA 51–58                   |
|                                 | YKELIGFQG 147–154                             | ISDAIIHV 108–115                                  | EAEMKASE 53–60                   |
|                                 |                                               | MTKAILFEL 132–139                                 | GGILIKKKG 74–81                  |
|                                 |                                               | ALELIRNDF 135–142                                 | EAEKLPFA 84–91                   |
|                                 |                                               | KYKELIGFQ 146–153                                 | PDDFGADA 121–128                 |
|                                 |                                               |                                                   | QGAMITKAL 129–136                |
|                                 |                                               |                                                   | GAMTIKALE 130–137                |
|                                 |                                               |                                                   | TKAILELFR 133–140                |
|                                 |                                               |                                                   | KALELRNF 134–141                 |
|                                 |                                               |                                                   | AAYKIEKFIA 144–151               |
|                                 |                                               |                                                   | ELGFQQC 149–154                  |
|                                 |                                               |                                                   |                                  |
| **Cocoa 21-kDa seed protein**   |                                               |                                                   |                                  |
| (SwissProt no. P32765)          |                                               |                                                   |                                  |
| GLAILGRA                        | GGLAILGRA 57–64                               | VANAANSP 23–30                                   | GRATIGGSC 62–69                  |
|                                 | GLAILGRAT 58–65                               | YYVUSIS 45–52                                     | CPEIVVQR 69–76                   |
|                                 | ATGQSCEP 64–71                                | EIVQQRS 71–78                                     | VRVSTIDVN 98–105                 |
|                                 | GKWVWTDD 132–139                              | VVQRRSD 72–79                                     | NIEFIVPR 105–112                 |
|                                 | GYKFPCP 163–170                               | PVIFSNAK 85–92                                    | PIRDRLCS 110–117                 |
|                                 | KRFICPSV 165–172                              | VIFSNADS 86–93                                    | TSTWWRDL 118–125                 |
|                                 |                                               | AGKWWVVT 131–138                                  | AGVLUYMKF 159–166                |
|                                 |                                               | PNTLCSWF 147–154                                  | SVCDSCFT 171–178                 |
|                                 |                                               | TLCWSWFK 149–156                                  | TSUWCIR 178–178                  |
|                                 |                                               | LCSWKFIE 150–157                                  | TSTWMODE 187–194                 |
|                                 |                                               | CSWFIEKIE 151–158                                 | IRLADSND 193–200                 |
|                                 |                                               | QIRLALSD 192–199                                  | RLALSDNE 194–201                 |
|                                 |                                               | ASKTIIKQV 209–216                                 |                                  |
|                                 |                                               |                                                   |                                  |
| **Cocoa vicilin**               |                                               |                                                   |                                  |
| (TrEMBL no. A0A061EM85)         |                                               |                                                   |                                  |
| NDYRLAMF                       | NDYRLAMF 50–57                               | PKRRISFQT 17–24                                   | RSEEDEEQ 1–8                     |
|                                 | ENKRESMNV 91–98                               | RRSFQTRF 19–26                                    | PYYPFKRKR 13–20                  |
|                                 | TYYVVSQD 111–118                              | EGNKFILQ 30–37                                    | YYFPPKRRS 14–21                  |
|                                 | GMFRKIkA 190–197                              | FKLQRFA 33–40                                     | YFPKRIKSF 15–22                  |
|                                 | KAKPEAQIR 194–201                             | LQRFANES 36–43                                    | RSFQTRFR 20–27                   |
|                                 | AKPEIRAQ 195–202                              | KGINIVYRL 47–54                                   | FQTRFRIDE 22–29                  |
|                                 | KPEQIRAI 196–203                              | GINDYRLA 48–55                                    | QTRFRIDEE 23–30                  |
|                                 | ERLAIINL 216–223                              | DRYLAMFEE 51–58                                   | KILQRFAE 34–41                   |
|                                 | FKLNNQGA 257–264                              | RLAMIFEEAN 53–60                                   | ILQRPAEN 35–42                   |
|                                 | VPHYINSKA 266–273                             | CDAEAIYF 70–77                                    | NPNFTILP 60–67                   |
|                                 | GYAMQMACP 284–291                             | EAIYFVTN 73–80                                    | DAEAIYFV 71–78                   |
|                                 | VTFIFA5D 343–350                              | TITFVTHE 84–91                                    | AEAHYFVT 72–79                   |
|                                 | LVNDNIFMN 395–402                             | TVSVPPAG 102–109                                  | AAYFVTNG 74–81                   |
|                                 |                                               | SVPAGSTV 105–112                                  | GTIIFVT 83–90                    |
|                                 |                                               | STVVVSQ 110–117                                   | VTHEIKES 88–95                   |
|                                 |                                               | TIAVLALP 124–131                                  | KESNVQQR 93–100                  |
|                                 |                                               | VLALPVS 127–134                                   | ESYNVQQR 94–101                  |
Table 1 (continued)

| Substrate | Cleavage sites specific for the cocoa protease | Common cleavage sites of cocoa protease and pepsin | Cleavage sites specific for pepsin |
|-----------|-----------------------------------------------|-------------------------------------------------|---------------------------------|
|           | P4–P4 Position                                | P4–P4 Position                                  | P4–P4 Position                  |
| KYEL/FFP  | 137–144                                       | YNVQ/RGT  | 96–103 |
| ELEF/PAGN | 139–146                                       | VRQG/TVVS | 98–105 |
| NKPE/SSYG | 147–154                                       | RGTV/VSVP | 100–107 |
| YGAS/SEYV | 153–160                                       | GTSV/SVPA | 101–108 |
| YEVE/LTFV | 158–165                                       | VSVP/PGS  | 103–110 |
| REK/EELE  | 169–176                                       | AGSTVVV  | 108–115 |
| KLE/ELEE  | 171–178                                       | GSTVVVS  | 109–116 |
| EKE/EQKR  | 173–180                                       | LTIAVLAL | 123–130 |
| QRAIS/SQ  | 199–206                                       | IAVALPV  | 125–132 |
| GEL/AINL  | 215–222                                       | PKYIELF  | 135–142 |
| AIN/LSQS  | 219–226                                       | GKYEIIFP | 136–143 |
| NGR/FYAC  | 233–240                                       | YELFPAG  | 138–145 |
| AVSA/FKLN | 253–260                                       | PESY/YGAF | 149–156 |
| NQGA/IFVP | 260–267                                       | YYGAISY  | 152–159 |
| KAD/VFV  | 272–279                                       | GAFSYEVL | 154–161 |
| SGR/DRRE  | 302–309                                       | AFSEYVE  | 155–162 |
| GRQ/DRRQ  | 303–310                                       | FSEYVE  | 156–163 |
| RQDR/REQE | 304–311                                       | EVLETFVN | 159–166 |
| EEPROM/E  | 316–323                                       | ETVF/NTQR | 162–169 |
| FGE/PQKV  | 319–326                                       | QGCMFRK  | 188–195 |
| GDVF/VAPA | 332–339                                       | LAIN/LSQ  | 218–225 |
| AVTF/SFK  | 342–349                                       | INLLSQS  | 220–227 |
| AVA/FGLN  | 355–362                                       | GRFFEACP | 234–241 |
| QRFI/LACK | 366–373                                       | FSQF/QMD | 244–251 |
| KKN/LVRQM | 373–380                                       | VSAFLK  | 254–261 |
| EAK/ELSF  | 383–390                                       | AFKLNQ  | 256–263 |
| FSK/LVDN  | 392–399                                       | GAIFVPH  | 262–269 |
| ESYS/FMS  | 405–412                                       | FVV/FVTD  | 275–282 |
|           |                                               | CPHEL/SRQS | 290–297 |
|           |                                               | SRQ/SQS/SQ | 294–301 |
|           |                                               | RQS/G/SQS | 295–302 |
|           |                                               | SQG/SQ/SQR | 297–304 |
|           |                                               | QGQ/QR/QG | 298–305 |
|           |                                               | GQS/QR/GQ | 299–306 |
|           |                                               | SQG/QRQR  | 300–307 |
|           |                                               | EEE/FGEF | 315–322 |
|           |                                               | PGD/VP/VAP | 331–338 |
|           |                                               | PLNA/AVFG | 352–359 |
|           |                                               | NAVA/FGLN | 354–361 |
|           |                                               | AFGLNAQ  | 357–364 |
|           |                                               | FGLNAQNN | 358–365 |
|           |                                               | NNQRIFLA | 364–371 |
|           |                                               | RIFL/AGKK | 367–374 |
|           |                                               | IFLAGGKN | 368–375 |
|           |                                               | FLAGIKKNL | 369–376 |
|           |                                               | VRQM/IDEA | 377–384 |
|           |                                               | RQM/DEAK | 378–385 |
|           |                                               | QMD/SAKE | 379–386 |
|           |                                               | MDSE/AKEL | 380–387 |
|           |                                               | GVP/SILVD | 390–397 |
|           |                                               | DNFIN/PND | 397–404 |
|           |                                               | NNP/DIESY | 401–408 |
|           |                                               | PDES/YFMS | 403–410 |
|           |                                               | SQQR/QRG | 412–419 |
|           |                                               | QQQR/QRGDE | 413–420 |

* Octapeptide sequences around the cleavage sites for the cocoa aspartic protease and pepsin, respectively, detected by partial proteolysis of myoglobin, the cocoa 21-kDa seed protein, and the cocoa vicilin-class(7S) globulin. Data were separately listed for sites exclusively cleaved by the cocoa aspartic protease and pepsin, respectively, and those cleaved by both proteases (= unspecific cleavage sites).
Table 2
Putative cleavage sites of the cocoa aspartic protease predicted from oligopeptides isolated from fermented cocoa beans.

| Substrate | Putative cleavage site | Position | N- or C-terminal localization of the cleavage site | Cleavage site also detected in vitro | References |
|-----------|------------------------|----------|--------------------------------------------------|-------------------------------------|------------|
| **Cocoa 21-kDa seed protein** (SwissProt no. P32765) | | | | | |
| VANAANSP | 23–30 | N-terminal | yes | | [1] |
| SPVLDTDG | 29–36 | C-terminal | no | | [3] |
| YYVLSSIS | 45–52 | N-terminal | yes | | [3] |
| SSISGAGG | 49–56 | N-terminal | no | | [3] |
| GGLIALGR | 56–63 | C-terminal | no | | [3] |
| IVVQRSD | 72–79 | N-terminal | yes | | [3] |
| SDLINGTP | 78–85 | N-terminal | no | | [3] |
| PVIFSNAD | 85–92 | N- and C-terminal | no | | [3] |
| FSNADS | 88–95 | N-terminal | no | | [3] |
| DVVRVSTD | 96–103 | N-terminal | no | | [3] |
| TDVNLIEFV | 102–109 | N- and C-terminal | no | | [3] |
| NIEFVIPR | 105–112 | C-terminal | no | | [3] |
| CSTSTTVWR | 116–123 | N-terminal | no | | [3] |
| STVWRLDN | 119–126 | N-terminal | no | | [3] |
| WRLDNYDN | 122–129 | C-terminal | no | | [3] |
| LALSIDNEW | 195–202 | N-terminal | no | | [3] |
| AWMFIKKAS | 203–210 | C-terminal | no | | [3] |
| **Cocoa vicilin** (TrEMBL no. A0A061EM85) | | | | | |
| EGQQRNNP | 6–13 | N- and C-terminal | no | | [3,4] |
| GQQRNPPY | 7–14 | N-terminal | no | | [3,4] |
| QQRNPPYY | 8–15 | N-terminal | no | | [4] |
| QRNNPYYF | 9–16 | N-terminal | no | | [4] |
| PYYFIPKR | 13–20 | C-terminal+CP | no | | [4] |
| YFPKRRSF | 15–22 | N- and C-terminal | no | | [3,4] |
| FKKRIRSFQ | 16–23 | N-terminal | no | | [4] |
| RRSFPOTR | 19–26 | C-terminal | yes | | [3,4] |
| RSPQCFR | 20–27 | N-terminal | no | | [3] |
| TRFRDIEG | 24–31 | N-terminal | no | | [3] |
| RDEEICNF | 27–34 | N- and C-terminal | no | | [3,4] |
| EEENFKNL | 29–36 | N-terminal | no | | [3] |
| ECNFIXLQ | 30–37 | N- and C-terminal | yes | | [3,4] |
| FKILQRFA | 33–40 | C-terminal | yes | | [3] |
| KILQRFAE | 34–41 | C-terminal | no | | [4] |
| SPPLKGIN | 43–50 | N-terminal | no | | [4] |
| KGIDRVR | 47–54 | C-terminal | yes | | [4] |
| INDYIRAM | 49–56 | N-terminal | no | | [4] |
| RLAMFEEAN | 53–60 | C-terminal+CP | yes | | [4] |
| NPNIFILP | 60–67 | N-terminal | no | | [4] |
| IILPHICDA | 65–72 | C-terminal | no | | [4] |
| YFVTINGK | 76–83 | N-terminal | no | | [3] |
| VTNKGTG | 78–85 | N-terminal | no | | [4] |
| TTIFVFTH | 84–91 | C-terminal + CP | yes | | [3,4] |
| THENKESY | 89–95 | N-terminal | no | | [3] |
| YNQVRQGT | 96–103 | N- and C-terminal | no | | [3,4] |
| TVVSVPAG | 102–109 | C-terminal | yes | | [4] |
| VLAIVPNS | 127–134 | N-terminal | yes | | [4] |
| LPVNSPGK | 129–138 | N-terminal | no | | [4] |
| PGCYIEFF | 135–142 | C-terminal | no | | [4] |
| FPAGNNK | 142–149 | N-terminal | no | | [3] |
| AGNNK | 144–151 | N-terminal | no | | [4] |
| NKPESSYY | 147–154 | C-terminal | no | | [3] |
| KPESSSYY | 148–155 | N- and C-terminal | no | | [3,4] |
| FSYEVLET | 156–163 | N-terminal | no | | [3] |
| YEVLETCV | 158–167 | C-terminal | yes | | [3] |
| EVLETVP | 159–166 | C-terminal | no | | [3] |
| PRHRGRAY | 209–217 | N-terminal | no | | [4] |
Cocoa protease, the cocoa 21-kDa seed protein, and the cocoa vicilin-class (7S) globular storage protein were isolated from the acetone-dry powder of unfermented cocoa beans essentially as described.

| Substrate | Putative cleavage site | Position | N- or C-terminal localization of the cleavage site | Cleavage site also detected in vitro [1] | References |
|-----------|------------------------|----------|--------------------------------------------------|----------------------------------------|------------|
| ERLLINL | 216–223 | N-terminal | yes | [4] |
| AINLISQS | 219–226 | C-terminal + CP | yes | [4] |
| INLISQSP | 220–227 | C-terminal | no | [4] |
| NLLSISQP | 221–228 | C-terminal | no | [4] |
| VAVSAFKL | 252–259 | N-terminal | no | [4] |
| AVSAFKLN | 253–260 | N-terminal | yes | [4] |
| FKLNQGAIF | 257–264 | C-terminal + CP | yes | [4] |
| LNOQGAIF | 258–265 | N- and C-terminal | no | [4] |
| LNQGAI | 259–266 | N-terminal | no | [4] |
| NQGAIFV | 260–267 | N- and C-terminal | yes | [4] |
| QGAI | 262–269 | N-terminal | no | [4] |
| GAIVPH | 264–273 | C-terminal + CP | yes | [4] |
| VPHYN | 266–274 | C-terminal | no | [4] |
| HYNSIKAT | 268–275 | C-terminal | no | [4] |
| KATF | 272–279 | C-terminal + CP | yes | [4] |
| SQSGIKQDR | 300–307 | N-terminal | no | [3] |
| EQEEI | 309–316 | C-terminal | no | [3] |
| GEFQIKQA | 320–327 | N-terminal | no | [4] |
| QQYKAPLS | 323–330 | N-terminal | yes | [4] |
| KAPLSPGD | 326–333 | N- and C-terminal | no | [3, 4] |
| APLSIPDVF | 327–334 | N-terminal | no | [3] |
| GDVFVAPA | 332–339 | N- and C-terminal | yes | [3, 4] |
| VFVAI | 334–341 | N-terminal | no | [3] |
| APAGHA | 337–344 | N-terminal | no | [4] |
| AVTF | 342–349 | C-terminal | yes | [3, 4] |
| VTFIASKD | 343–350 | N- and C-terminal | yes | [3, 4] |
| FFSDKQP | 345–352 | N-terminal | no | [3] |
| FASKDQPL | 346–353 | N-terminal | no | [4] |
| AVAFGLNA | 355–362 | C-terminal + CP | yes | [3, 4] |
| LNAQNNQR | 360–367 | N-terminal | no | [4] |
| NAQNNQRI | 361–368 | N-terminal | no | [4] |
| AQQNNQRF | 362–369 | N-terminal | no | [4] |
| QNNQIRIF | 363–370 | N-terminal | no | [4] |
| QRIFLAGK | 366–373 | C-terminal | no | [4] |
| GKKNIVRQ | 372–379 | N-terminal | no | [4] |
| NLVRQMDS | 375–382 | C-terminal | no | [4] |
| AKELISF | 384–391 | N-terminal | no | [4] |
| KELSFGVP | 385–392 | N-terminal | no | [4] |
| PSKLVNDI | 392–399 | C-terminal + CP | no | [4] |
| NPDIESYF | 402–409 | N-terminal | no | [4] |
| ESYF | 405–412 | C-terminal | no | [4] |

* Octapeptide sequence ([P4–P4]) around the putative cleavage site.

* Position of the octapeptide in the amino acid sequence of the degraded seed protein.

* Localization of the cleavage site at the N-terminal or C-terminal end of the oligopeptide, from which the cleavage site was predicted. Since the peptides formed during cocoa fermentation are modified by a carboxypeptidase [2, 5], the N-terminal cleavage sites are more reliable than the C-terminal ones. In case of the C-terminal ends of the corresponding oligopeptide, a downstream localized cleavage site was predicted, whenever the resulting peptide fragment could be modified by the cocoa carboxypeptidase [6] to the finally detected oligopeptide (indicated by “+ CP”).

2. Experimental design, materials and methods

2.1. Determination of cleavage sites by in-vitro proteolysis

Cocoa protease, the cocoa 21-kDa seed protein, and the cocoa vicilin-class (7S) globular storage protein were isolated from the acetone-dry powder of unfermented cocoa beans essentially as described.
10 mg of horse myoglobin or of the individual cocoa seed proteins in 1 ml of 20 mM sodium acetate (pH 5.0) were partially digested with either 100 μg of purified cocoa aspartic protease or 50 μg of commercial porcine pepsin (Sigma-Aldrich Chemie, Taufkirchen, Germany). The obtained peptides were modified by reduction with dithiothreitol and subsequent alkylation of the cysteine residues with iodoacetamide before being analyzed by mass spectrometry.

Table 3
Abundance of different amino acid residues in the P4 to P4 -i positions of the predicted and experimentally detected cleavage sites of the cocoa aspartic protease.

|     | P4 | P3 | P2 | P1 |
|-----|----|----|----|----|
|     | In-situ | In-vitro | In-situ | In-vitro | In-situ | In-vitro | In-situ | In-vitro | In-situ | In-vitro |
| W   | 0.93 | 0.93  | 0.93  | 0.93  | 1.02  | 1.88   | 1.02  | 1.88   | 1.02  | 1.88   |
| F   | 6.54 | 1.88  | 1.88  | 1.88  | 0.00  | 1.88   | 1.88  | 1.88   | 1.88  | 1.88   |
| Y   | 7.14 | 6.54  | 11.32 | 5.54  | 6.12  | 6.54   | 6.12  | 6.54   | 21.24 | 20.56  |
| L   | 7.14 | 5.54  | 8.41  | 5.54  | 11.32 | 5.54   | 11.32 | 5.54   | 21.24 | 20.56  |
| I   | 0.00 | 0.93  | 0.93  | 0.93  | 1.02  | 0.00   | 1.02  | 0.00   | 1.02  | 0.00   |
| M   | 7.14 | 5.54  | 8.16  | 10.29 | 16.32 | 11.32  | 16.32 | 11.32  | 21.24 | 20.56  |
| V   | 10.28 | 7.14  | 9.34  | 3.06  | 3.06  | 3.76   | 3.06  | 3.76   | 9.18  | 5.54   |
| A   | 7.14 | 11.20 | 4.08  | 5.54  | 12.24 | 4.67   | 13.26 | 2.80   | 6.12  | 10.28  |
| D   | 0.00 | 1.88  | 4.08  | 4.67  | 2.04  | 2.80   | 2.04  | 2.80   | 6.12  | 3.76   |
| H   | 0.00 | 2.80  | 4.08  | 4.67  | 2.04  | 1.88   | 2.04  | 1.88   | 6.12  | 3.76   |
| R   | 0.00 | 3.74  | 7.14  | 4.67  | 10.24 | 9.34   | 10.24 | 9.34   | 6.12  | 3.76   |
| K   | 7.14 | 9.34  | 5.10  | 9.34  | 4.08  | 11.32  | 4.08  | 11.32  | 9.18  | 5.54   |
| P   | 7.14 | 3.74  | 11.22 | 2.80  | 5.10  | 2.80   | 1.02  | 0.93   | 6.12  | 3.76   |

a Amino acid positions around the cleavage sites.
b Predicted from the N-terminal and C-terminal ends of oligopeptides isolated from fermented cocoa beans [3,4].
c Detected by in vitro digestion of three different protein substrates with the cocoa aspartic protease (compare Table 1).
d Values are expressed in percent of all amino acids found in these positions. Values above 6% are marked in bold.
Liquid chromatography-MALDI-TOF/TOF-MS/MS analyses were performed on a 4700 proteomics Analyzer (ABSCIEX, Framingham, MS) off-line coupled with an Ultimate HPLC system and Probot fractionation device (both Dionex/Thermo, Idstein, Germany). LC separations were performed on an analytical column (PepMap C18, 3 μm, 150 mm × 75 μm; Dionex) at a flow rate of 200 nl/min. Mobile phase (A) was 2:98 (v/v) acetonitrile/water containing 0.05% (v/v) TFA and (B) was 80:20 (v/v) acetonitrile/water containing 0.045% (v/v) TFA. Gradients were 0–10% B in 4 min, 10-50% B in 30 min, 50–100% B in 2 min. Column effluent was continuously mixed with MALDI matrix (5 mg/ml α-cyano-4-hydroxycinnamic acid in 70:30 (v/v) acetonitrile/water containing 0.1% (v/v) TFA, 1 μl/min) and spotted at 10-s intervals on 26 × 12 spot arrays on MALDI steel targets (Applied Biosystems, Darmstadt, Germany).

Mass spectra were acquired in a data-dependent mode. The MS spectra were recorded in the mass range of m/z 800–4000 and with the accumulation of 2000 subspectra. MS/MS spectra were measured from the five most intensive precursor ions (S/N > 30). 5000–10,000 laser shots were accumulated. MS and MS/MS peak lists were generated by the “Peak to Mascot” tool of the 4000er Series Explorer v3.6. For MS/MS data analysis, MASCOT server (version 2.3, Matrixscience, London, UK) was used. Data base searches were performed using SwissProt (2015_03; 547964 protein sequences) and the following parameters: no enzyme, one missed cleavage, variable modifications: carbamidomethylation (C), oxidation (M), pyro-glu (Q), mass tolerances for MS and MS/MS: 100 ppm and 0.3 Da. Enzymatic peptides of horse myoglobin (SwissProt no. P68082), cocoa vicilin-class(7S) globulin (TrEMBL no. A0A061EM85), and the cocoa 21-kDa seed protein (SwissProt no. P32765) were accepted as identified if their MS/MS spectra provided a MASCOT score for identity with p < 0.05.

The different cleavage sites were determined by localization of the N- and C-terminal ends of the oligopeptides within the amino acid sequence of the corresponding substrate proteins. The octapeptide sequences around the cleavage sites and their positions in the corresponding substrate proteins are listed in Table 1. Three classes of cleavage sites were found and separately listed (Table 1):

(1) Those which were exclusively cleaved by the cocoa aspartic protease (=specific cleavage sites of the cocoa enzyme),
(2) those which were cleaved both by the cocoa aspartic protease and pepsin (=unspecific cleavage sites of the cocoa enzyme) and
(3) those which were exclusively cleaved by pepsin.

2.2. Determination of putative in-situ cleavage sites used during cocoa fermentation

Oligopeptides isolated from fermented cocoa beans and sequenced by ESI-MS/MS mass spectrometric analyses were taken from the literature [3,4] and used to identify the putative in-situ cleavage sites of the cocoa aspartic protease in the 21-kDa cocoa seed protein and in the vicilin-class(7S) globulin of the cocoa beans, respectively. The octapeptide sequences around the putative cleavage sites used in the formation of the oligopeptides isolated from fermented cocoa beans and their positions in the amino acid sequences of the 21-kDa cocoa seed protein and the cocoa vicilin-class(7S) globulin, respectively, are listed in Table 2. Since the oligopeptides generated during fermentation of the cocoa beans are more or less modified at their C-terminal ends due to the activity of a carboxypeptidase [5], prediction of the C-terminal cleavage sites is less reliable than the cleavage sites predicted from the N-terminal ends. Due to the known cleavage specificity of this particular carboxypeptidase [6], however, the putative cleavage sites corresponding to the C-terminal ends of the original cleavage products generated by the cocoa aspartic protease can be predicted with at least some reliability. When the predicted C-terminal cleavage site was assumed to be downstream from the C-terminal end of the isolated peptide, this was marked by “+CP”. Up to now, 87 different oligopeptides have been isolated from fermented cocoa beans and sequenced by mass spectrometry [3,4]. All these oligopeptides were derived from the 21-kDa seed protein and the cocoa vicilin-class (7S) globulin, respectively [3,4].
From the N- and C-terminal ends of these 87 oligopeptides, 98 putative cleavage sites of the cocoa aspartic protease have been predicted (Table 2), 23 of which being identical to cleavage sites detected by in-vitro proteolysis (Tables 1 and 2).

To get an insight into the cleavage specificity of the cocoa aspartic protease, the relative abundance of the different amino acid residues in the P4–P4’ positions around the cleavage sites have been determined (Table 3). This was done both for the cleavage sites putatively used in-situ (during the fermentation process) and for the cleavage sites determined by in-vitro proteolysis (Table 3). In the latter case, all the cleavage sites of the cocoa aspartic protease have been considered, i.e. without discrimination between specific and unspecific cleavage sites as done in Table 1. Considerable differences have been observed for the relative abundance of some amino acids in the P4–P4’ positions between the in-situ (used during fermentation) and the in-vitro cleavage sites, respectively (Table 3). Analysis of chemical compounds by MALDI-TOF-MS used for the identification of peptide fragments generated during in-vitro proteolysis [1] is restricted to ions with m/z > 799, due to ions generated from the matrix components. As recently reported, most peptides present in fermented cocoa beans, however, have molecular masses below this limit [3,4]. Therefore, considerably more peptides and their corresponding N- and C-terminal ends can be detected and analyzed by LC-ESI-MS/MS than by LC-MALDI-TOF/TOF-MS/MS.

Transparency document. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.06.021.

References

[1] K. Janek, A. Niewienda, J. Wöstemeyer, J. Voigt, The cleavage specificity of the aspartic protease of cocoa beans involved in the generation of the cocoa-specific aroma precursors, Food Chem. 211 (2016) 320–328.
[2] J. Voigt, H. Heinrichs, G. Voigt, D. Wrann, B. Biehl, Cocoa-specific aroma precursors are generated by proteolytic digestion of the vicilin-like globulin of the cocoa seeds, Food Chem. 50 (1994) 177–184.
[3] A. Marseglia, S. Sforza, A. Faccini, M. Bencivenni, G. Palla, A. Caligiani, Extraction, identification and semi-quantification of oligopeptides in cocoa beans, Food Res. Intern. 63 (2014) 382–389.
[4] J. Voigt, K. Janek, K. Textoris-Taube, A. Niewienda, J. Wöstemeyer, Partial purification and characterisation of the peptide precursors of the cocoa-specific aroma components, Food Chem. 192 (2016) 706–713.
[5] J. Voigt, B. Biehl, H. Heinrichs, S. Kamaruddin, S. Gaim Marsoner, A. Hugi, In-vitro formation of cocoa-specific aroma precursors: aroma-related peptides generated from cocoa-seed protein by co-operation of an aspartic endoprotease and a carboxypeptidase, Food Chem. 49 (1994) 173–180.
[6] G. Bytof, B. Biehl, H. Heinrichs, J. Voigt, Specificity and stability of the carboxypeptidase activity in ripe, ungerminated seeds of Theobroma cacao L, Food Chem. 54 (1995) 15–21.