BMC Structural Biology

BioMed Central

Research article

Open Access

A model of tripeptidyl-peptidase I (CLN2), a ubiquitous and highly conserved member of the sedolisin family of serine-carboxyl peptidases

Alexander Wlodawer*[1], Stewart R Durell[2], Mi Li[1,3], Hiroshi Oyama[4], Kohei Oda[4] and Ben M Dunn[5]

Address: [1]Protein Structure Section, Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, MD 21702, USA, [2]Laboratory of Experimental and Computational Biology, National Cancer Institute, Bethesda, MD 20892, USA, [3]Basic Research Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, MD 21702, USA, [4]Department of Applied Biology, Faculty of Textile Science, Kyoto Institute of Technology, Sakyo-ku, Kyoto 606-8585, Japan and [5]Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, Florida 32610, USA

Email: Alexander Wlodawer* - wlodawer@ncifcrf.gov; Stewart R Durell - durells@exchange.nih.gov; Mi Li - mili@ncifcrf.gov; Hiroshi Oyama - oyama@ipc.kit.ac.jp; Kohei Oda - bika@ipc.kit.ac.jp; Ben M Dunn - bdunn@college.med.uf.edu

* Corresponding author

Published: 11 November 2003

BMC Structural Biology 2003, **3**:8

Received: 23 July 2003
Accepted: 11 November 2003

This article is available from: http://www.biomedcentral.com/1472-6807/3/8

Abstract

Background: Tripeptidyl-peptidase I, also known as CLN2, is a member of the family of sedolisins (serine-carboxyl peptidases). In humans, defects in expression of this enzyme lead to a fatal neurodegenerative disease, classical late-infantile neuronal ceroid lipofuscinosis. Similar enzymes have been found in the genomic sequences of several species, but neither systematic analyses of their distribution nor modeling of their structures have been previously attempted.

Results: We have analyzed the presence of orthologs of human CLN2 in the genomic sequences of a number of eukaryotic species. Enzymes with sequences sharing over 80% identity have been found in the genomes of macaque, mouse, rat, dog, and cow. Closely related, although clearly distinct, enzymes are present in fish (fugu and zebra), as well as in frogs (Xenopus tropicalis). A three-dimensional model of human CLN2 was built based mainly on the homology with *Pseudomonas* sp. 101 sedolisin.

Conclusion: CLN2 is very highly conserved and widely distributed among higher organisms and may play an important role in their life cycles. The model presented here indicates a very open and accessible active site that is almost completely conserved among all known CLN2 enzymes. This result is somehow surprising for a tripeptidase where the presence of a more constrained binding pocket was anticipated. This structural model should be useful in the search for the physiological substrates of these enzymes and in the design of more specific inhibitors of CLN2.

Background

Although the existence of tripeptidyl-peptidase I (TPP-I) was first noted over 40 years ago [1], the structural and mechanistic basis of its activity has been largely misunderstood until quite recently. The situation changed after it was shown that TPP-I is identical to an independently characterized enzyme named CLN2. It was also demonstrated that mutations leading to abolishment of the enzymatic activity of CLN2 were the direct cause of a fatal inherited neurodegenerative disease, classical late-infantile neuronal ceroid lipofuscinosis [2]. This important observation was followed by the identification of CLN2 as a serine peptidase [3,4], without, however, specifying its structural fold and the details of the catalytic site. More accurate placement of CLN2 within the context of a family of related enzymes became possible only after high-resolution crystal structures of two bacterial enzymes with a limited sequence similarity to CLN2, sedolisin and kumamolisin, became available [5–7]. These structures defined a novel family of enzymes, now called sedolisins or serine-carboxyl peptidases, that is characterized by the utilization of a fully conserved catalytic triad (Ser, Glu, Asp) and by the presence of an Asp in the oxyanion hole [8]. Sedolisin and its several variants (e.g., kumamolisin, aorsin [9], and physarolisin [10]) have now been found in archaea, bacteria, fungi and amoebae, whereas the higher organisms seem to contain only variants of CLN2 [8]. The physiological role of sedolisins in the lower organisms has not yet been elucidated.

Despite the potential medical importance of CLN2 and related enzymes, no systematic studies of their genomic distribution have been published to date. There are also no published reports of the crystallization of this enzyme. In the absence of an experimental structure obtained by crystallography or NMR it is sometimes necessary to resort to molecular modeling in order to provide a structural basis for the explanation of the biological properties of an enzyme, and, in particular, to initiate design of its inhibitors. Examples of such very successful and useful modeling efforts are provided by HIV protease [11], or very recently by the peptidase from a coronavirus involved in the severe acute respiratory syndrome [12], among others. We have now applied the tools of molecular homology modeling to predicting a structure of CLN2 that could be used as a basis for a search for the biological substrates of this family of enzymes and for the design of specific inhibitors.

Results and discussion

Sequence comparisons of the CLN2-like enzymes

Mammalian enzymes homologous to human CLN2 [2,4] form a subfamily of sedolisins with highly conserved sequences (Figure 1). These enzymes are expressed with a prosegment consisting of 195 residues that is cleaved off during maturation, yielding the active catalytic domain. Complete sequences are available for CLN2 from six species in which it has been found so far (human, macaque, dog, mouse, rat, and cow). The full-length enzymes consist of 563 amino acids arranged in a single polypeptide chain containing both the prosegment and the catalytic domain, with the exception of mouse CLN2 that has a single deletion in the prosegment. The overall sequence identity for these enzymes is 81%, whereas the similarity is 92%. A pairwise comparison of the human and mouse enzymes yields 88% identity and 94% similarity, considerably higher than the median 78.5% identity reported for all identified mouse-human orthologs [13]. Thus, mammalian CLN2 appears to be a highly conserved enzyme.

In addition to the mammals, CLN2-like enzymes are also found in fugu (puffer fish – unannotated record SINFRUP00000077297 in the NCBI fugu sequence database, http://www.ncbi.nlm.nih.gov/BLAST/Genome/fugu.html) and zebrafish (contig wz4596.2 in the zebrafish EST database, http://fisher.wustl.edu/fish_lab). Only a fragment of the sequence of the latter enzyme agreed with the former. However, a few comparatively minor modifications can bring the zebrafish sequence into good agreement with that of the fugu CLN2. These modifications include a deletion of a single nucleotide from a run of three, as well as three insertions of repeated nucleotide pairs (Figure 2). It must be stressed that these modifications are speculative and may lead to prediction of several incorrect amino acids; however, they bring the two sequences into good global agreement (69% identities and 83% similarities).

The available amino acid sequence of the fugu CLN2 analog, named by us sedolisin-TPP [8], is also in good agreement with the sequences of the mammalian orthologs (Figure 3). The only major difference in the translated amino acid sequence compared to the mammalian and zebrafish enzymes is in the amino terminus of the propeptide region that is shorter by 30 amino acids (not shown). It is very likely that this represents a fault in the assembled sequence rather than a real variation, since the current coding frame is not initiated with a methionine, and a few extra residues are present in the full genomic sequence available from the fugu sequencing consortium http://genome.jgi-psf.org/fugu6/fugu6.home.html.

CLN2 is present not only in fish, but also in amphibians, in particular in Xenopus tropicalis (a species of frog). A partial sequence of its sedolisin-TPP (AL594774) found in the EST database http://www.sanger.ac.uk/Projects/X_tropicalis/blast_server.shtml spans the middle part of the catalytic domain, without reaching the part of the active site closer to the N terminus that contains the aspartic and glutamic acids that belong to the catalytic triad. However, the sequenced part of the enzyme shows 75% identity with the fugu sedolisin-TPP, and 69% identity with human CLN2 (Figure 3). Sequence similarity to bacterial or fungal sedolisins is much lower, indicating that the enzyme found in frogs might also share the functional properties of the CLN2 subfamily.

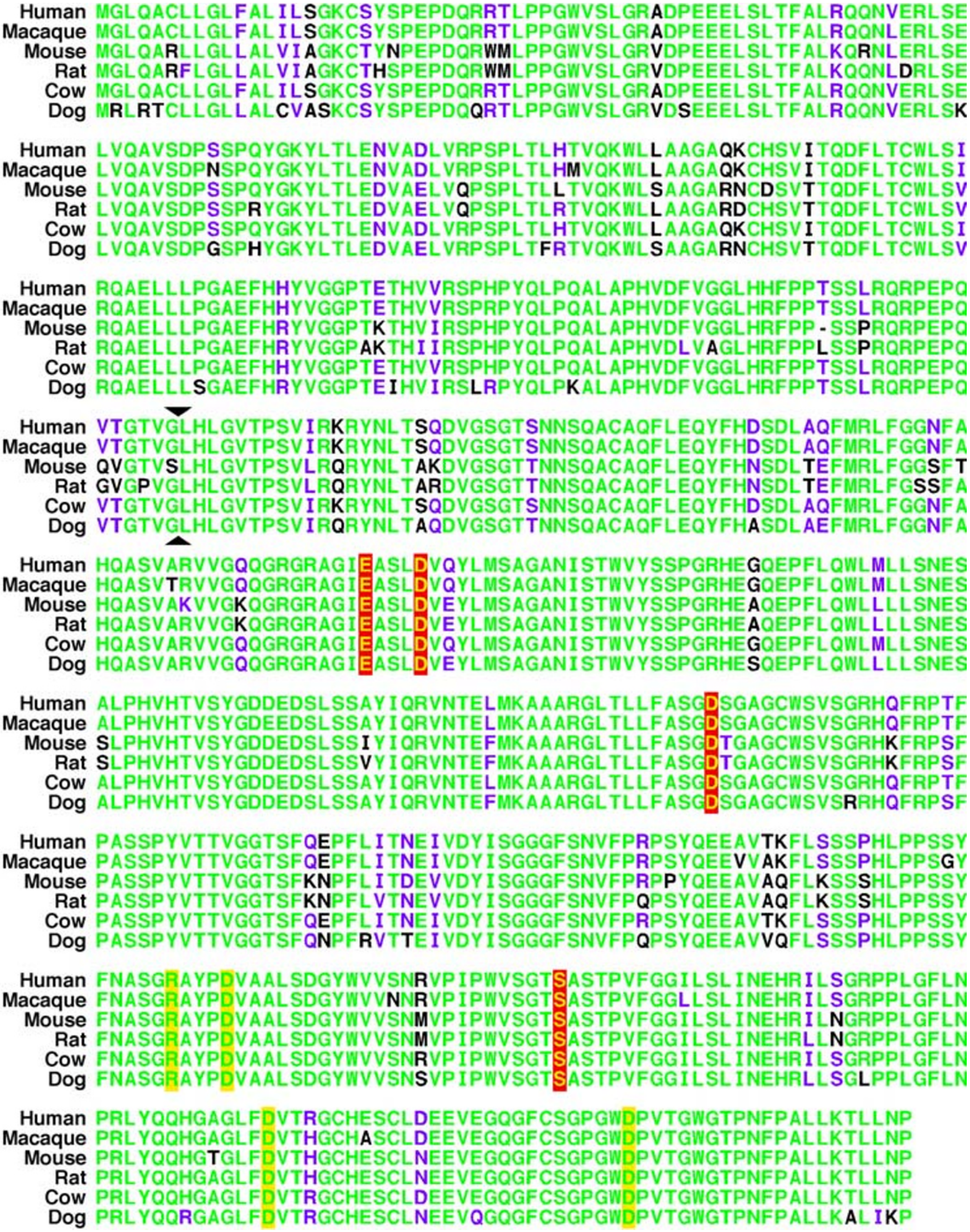
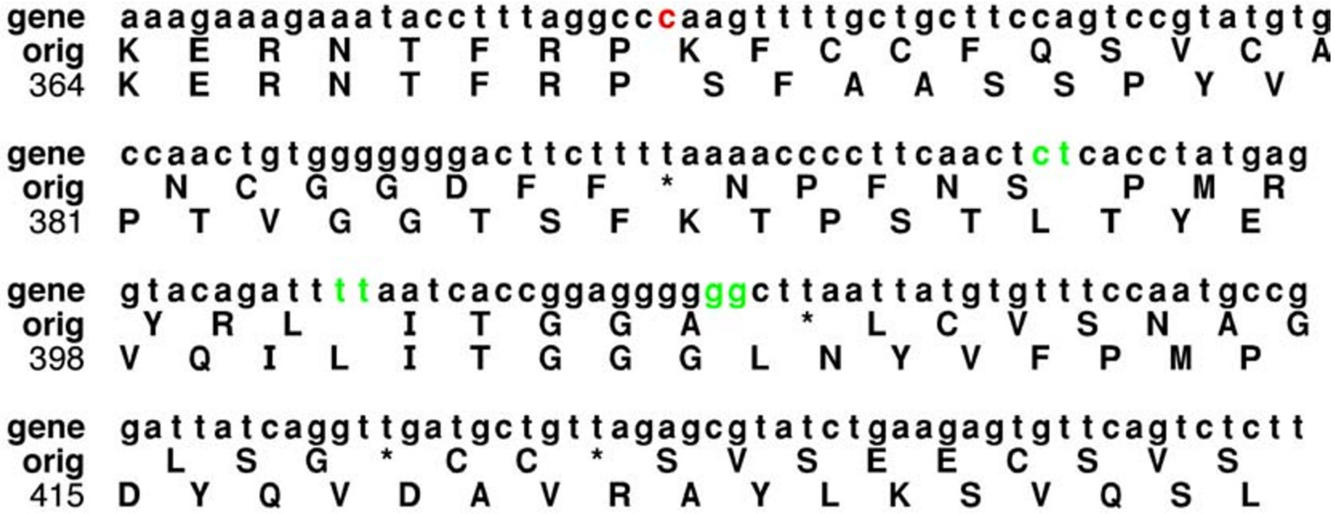

Human MGLQACLLGLFALILSGKCSYSPEPDQRRTLPPGWVSLGRADPEEELSLTFALRQQNVERLSE
Macaque MGLQACLLGLFALILSGKCSYSPEPDQRRTLPPGWVSLGRADPEEELSLTFALRQQNLERLSE
Mouse MGLQARLLGLLALVIAGKCTYNPEPDQRWMLPPGWVSLGRVDPEEELSLTFALKQRNLERLSE
Rat MGLQARFLGLLALVIAGKCTHSPEPDQRWMLPPGWVSLGRVDPEEELSLTFALKQQNLDRLSE
Cow MGLQACLLGLFALILSGKCSYSPEPDQRRTLPPGWVSLGRADPEEELSLTFALRQQNVERLSE
Dog MRLRTCLLGLLALCVASKCSYSPEPDQQRTLPPGWVSLGRVDSEEELSLTFALRQQNVERLSK
Human LVQAVSDPSSPQYGKYLTLENVADLVRPSPLTLHTVQKWLLAAGAQKCHSVITQDFLTCWLSI
Macaque LVQAVSDPNSPQYGKYLTLENVADLVRPSPLTLHMVQKWLLAAGAQKCHSVITQDFLTCWLSI
Mouse LVQAVSDPSSPQYGKYLTLEDVAELVQPSPLTLLTVQKWLSAAGARNCDSVTTQDFLTCWLSV
Rat LVQAVSDPSSPRYGKYLTLEDVAELVQPSPLTLRTVQKWLLAAGARDCHSVTTQDFLTCWLSV
Cow LVQAVSDPSSPQYGKYLTLENVADLVRPSPLTLHTVQKWLLAAGAQKCHSVITQDFLTCWLSI
Dog LVQAVSDPGSPHYGKYLTLEDVAELVRPSPLTFRTVQKWLSAAGARNCHSVTTQDFLTCWLSV
Human RQAELLLPGAEFHHYVGGPTETHVVRSPHPYQLPQALAPHVDFVGGLHHFPPTSSLRQRPEPQ
Macaque RQAELLLPGAEFHHYVGGPTETHVVRSPRPYQLPQALAPHVDFVGGLHRFPPTSSLRQRPEPQ
Mouse RQAELLLPGAEFHRYVGGPTKTHVIRSPHPYQLPQALAPHVDFVGGLHRFPP-SSPRQRPEPQ
Rat RQAELLLPGAEFHRYVGGPAKTHIIRSPHPYQLPQALAPHVDLVAGLHRFPPLSSPRQRPEPQ
Cow RQAELLLPGAEFHHYVGGPTETHVVRSPHPYQLPQALAPHVDFVGGLHRFPPTSSLRQRPEPQ
Dog RQAELLLSGAEFHRYVGGPTEIHVIRSLRPYQLPKALAPHVDFVGGLHRFPPTSSLRQRPEPQ
Human VTGTVGLHLGVTPSVIRKRYNLTSQDVGSGTSNNSQACAQFLEQYFHDSDLAQFMRLFGGNFA
Macaque VTGTVGLHLGVTPSVIRKRYNLTSQDVGSGTSNNSQACAQFLEQYFHDSDLAQFMRLFGGNFA
Mouse QVGTVSLHLGVTPSVLRQRYNLTAKDVGSGTTNNSQACAQFLEQYFHNSDLTEFMRLFGGSFT
Rat GVGPVGLHLGVTPSVLRQRYNLTARDVGSGTTNNSQACAQFLEQYFHNSDLTEFMRLFGSSFA
Cow VTGTVGLHLGVTPSVIRKRYNLTSQDVGSGTSNNSQACAQFLEQYFHDSDLAQFMRLFGGNFA
Dog VTGTVGLHLGVTPSVIRQRYNLTAQDVGSGTTNNSQACAQFLEQYFHASDLAEFMRLFGGNFA
Human HQASVARVVGQQGRGRAGIEASLDVQYLMSAGANISTWVYSSPGRHEGQEPFLQWLMLLSNES
Macaque HQASVTRVVGQQGRGRAGIEASLDVQYLMSAGANISTWVYSSPGRHEGQEPFLQWLMLLSNES
Mouse HQASVAKVVGKQGRGRAGIEASLDVEYLMSAGANISTWVYSSPGRHEAQEPFLQWLLLLSNES
Rat HQASVARVVGKQGRGRAGIEASLDVEYLMSAGANISTWVYSSPGRHEAQEPFLQWLLLLSNES
Cow HQASVARVVGQQGRGRAGIEASLDVQYLMSAGANISTWVYSSPGRHEGQEPFLQWLMLLSNES
Dog HQASVARVVGQQGRGRAGIEASLDVEYLMSAGANISTWVYSSPGRHESQEPFLQWLLLLSNES
Human ALPHVHTVSYGDDEDSLSSAYIQRVNTELMKAAARGLTLLFASGDSGAGCWSVSGRHQFRPTF
Macaque ALPHVHTVSYGDEEDSLSSAYIQRVNTELMKAAARGLTLLFASGDSGAGCWSVSGRHQFRPTF
Mouse SLPHVHTVSYGDDEDSLSSIYIQRVNTEFMKAAARGLTLLFASGDTGAGCWSVSGRHKFRPSF
Rat SLPHVHTVSYGDDEDSLSSVYIQRVNTEFMKAAARGLTLLFASGDTGAGCWSVSGRHKFRPSF
Cow ALPHVHTVSYGDDEDSLSSAYIQRVNTELMKAAARGLTLLFASGDSGAGCWSVSGRHQFRPTF
Dog ALPHVHTVSYGDDEDSLSSAYIQRVNTEFMKAAARGLTLLFASGDSGAGCWSVSRRHQFRPSF
Human PASSPYVTTVGGTSFQEPFLITNEIVDYISGGGFSNVFPRPSYQEEAVTKFLSSSPHLPPSSY
Macaque PASSPYVTTVGGTSFQEPFLITNEIVDYISGGGFSNVFPRPSYQEEVVAKFLSSSPHLPPSGY
Mouse PASSPYVTTVGGTSFKNPFLITDEVVDYISGGGFSNVFPRPPYQEEAVAQFLKSSSHLPPSSY
Rat PASSPYVTTVGGTSFKNPFLVTNEVVDYISGGGFSNVFPQPSYQEEAVAQFLKSSSHLPPSSY
Cow PASSPYVTTVGGTSFQEPFLITNEIVDYISGGGFSNVFPRPSYQEEAVTKFLSSSPHLPPSSY
Dog PASSPYVTTVGGTSFQNPFRVTTEIVDYISGGGFSNVFPQPSYQEEAVVQFLSSSPHLPPSSY
Human FNASGRAYPDVAALSDGYWVVSNRVPIPWVSGTSASTPVFGGILSLINEHRILSGRPPLGFLN
Macaque FNASGRAYPDVAALSDGYWVVNNRVPIPWVSGTSASTPVFGGLLSLINEHRILSGRPPLGFLN
Mouse FNASGRAYPDVAALSDGYWVVSNMVPIPWVSGTSASTPVFGGILSLINEHRILNGRPPLGFLN
Rat FNASGRAYPDVAALSDGYWVVSNMVPIPWVSGTSASTPVFGGILSLINEHRLLNGRPPLGFLN
Cow FNASGRAYPDVAALSDGYWVVSNRVPIPWVSGTSASTPVFGGILSLINEHRILSGRPPLGFLN
Dog FNASGRAYPDVAALSDGYWVVSNSVPIPWVSGTSASTPVFGGILSLINEHRLLSGLPPLGFLN
Human PRLYQQHGAGLFDVTRGCHESCLDEEVEGQGFCSGPGWDPVTGWGTPNFPALLKTLLNP
Macaque PRLYQQHGAGLFDVTHGCHASCLDEEVEGQGFCSGPGWDPVTGWGTPNFPALLKTLLNP
Mouse PRLYQQHGTGLFDVTHGCHESCLNEEVEGQGFCSGPGWDPVTGWGTPNFPALLKTLLNP
Rat PRLYQQHGAGLFDVTHGCHESCLNEEVEGQGFCSGPGWDPVTGWGTPNFPALLKTLLNP
Cow PRLYQQHGAGLFDVTRGCHESCLDEEVEGQGFCSGPGWDPVTGWGTPNFPALLKTLLNP
Dog PRLYQQRGAGLFDVTRGCHESCLNEEVQGQGFCSGPGWDPVTGWGTPNFPALLKALIKP

Figure 1

Sequence comparisons of mammalian CLN2-like enzymes. These sequences correspond to the complete enzymes, including the prosegment. Residues forming the active site are shown in yellow on red background, other conserved residues identified as important for the stability of the enzyme are marked with yellow background, residues identical in at least 5 of the structures are green, and residues similar in their character are shown in magenta. The maturation cleavage point generating the N terminus of the active enzyme is marked with black triangles.



```
gene  aaagaaagaaatacctttaggcccaagttttgctgcttccagtccgtatgtg
orig  K  E  R  N  T  F  R  P  K  F  C  C  F  Q  S  V  C  A
 364  K  E  R  N  T  F  R  P   S  F  A  A  S  S  P  Y  V

gene  ccaactgtggggggacttcttttaaaaccccttcaactctcacctatgag
orig    N  C  G  G  D  F  F  *  N  P  F  N  S    P  M  R
 381  P  T  V  G  G  T  S  F  K  T  P  S  T  L  T  Y  E

gene  gtacagattttaatcaccggaggggggcttaattatgtgtttccaatgccg
orig   Y  R  L   I  T  G  G  A   *  L  C  V  S  N  A  G
 398  V  Q  I  L  I  T  G  G  G  L  N  Y  V  F  P  M  P

gene  gattatcaggttgatgctgttagagcgtatctgaagagtgttcagtctctt
orig    L  S  G  *  C  C  *  S  V  S  E  E  C  S  V  S
 415  D  Y  Q  V  D  A  V  R  A  Y  L  K  S  V  Q  S  L
```

Figure 2

Corrected gene sequence of the zebrafish CLN2. This putative sequence shows the manual corrections that bring it into alignment with the sequence of the fugu enzyme. Inserted nucleotides are marked in green and a deleted one in red.

The discovery of highly conserved CLN2-like enzymes not only in mammals but also in two fish species and one of frogs may indicate that these peptidases are universally present in the vertebrates, and that their important role identified in humans [2] and mice [14] might be a more general feature.

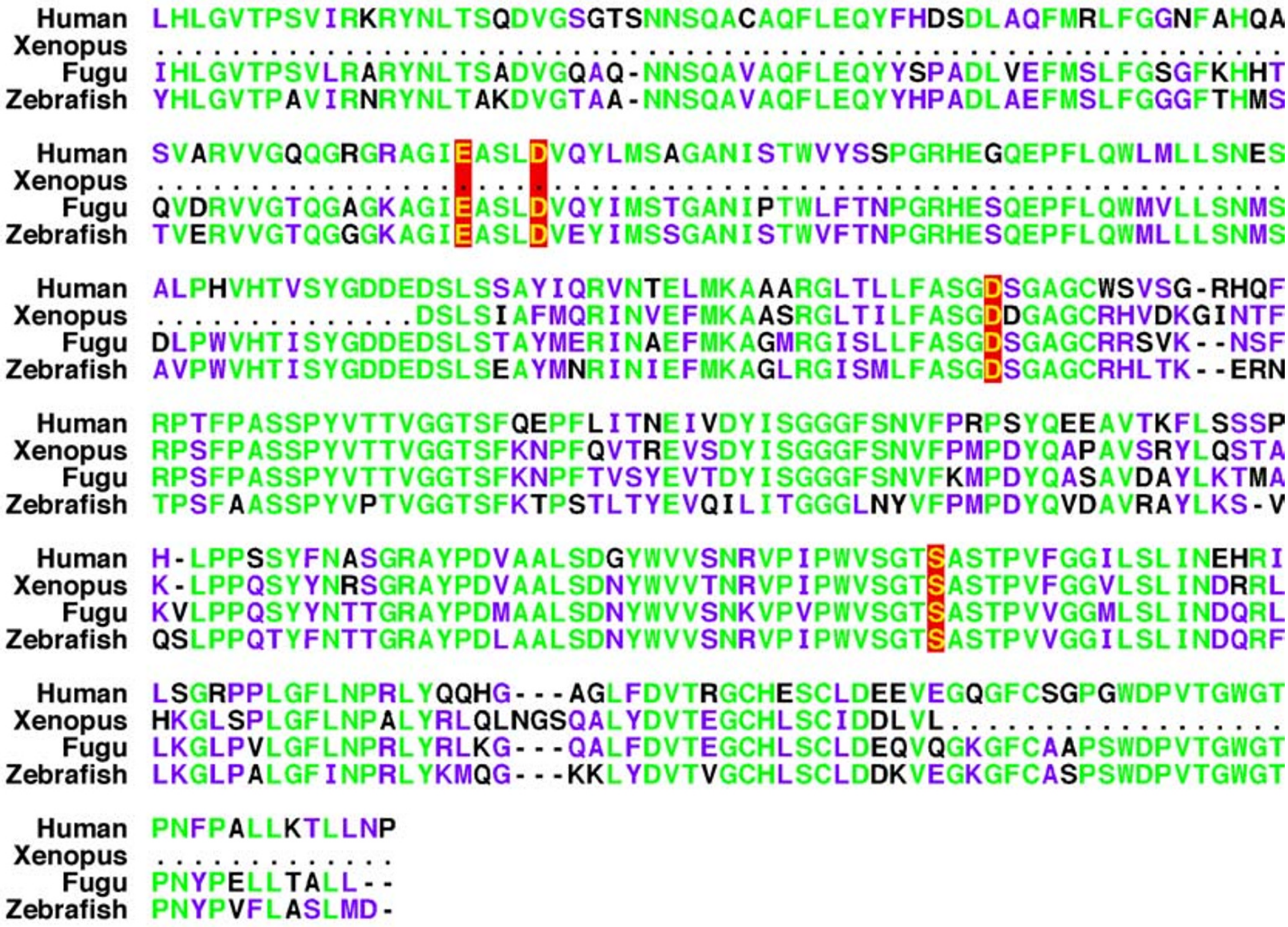
Modeling of the structure of human CLN2

The medical importance of CLN2 and the lack of a crystal structure inspired attempts at protein modeling. The first such model assumed that this enzyme is membrane-bound, with the sequence 271–294 (numbering corresponding to the mature enzyme, Figure 1) forming the putative transmembrane anchor [15]. However, in view of the subsequently obtained structures of the fully water-soluble sedolisins, this model was clearly incorrect. Exploiting the sequence similarity between CLN2, sedolisin, and kumamolisin (Figure 4), we have now used the experimentally obtained structures of the latter two enzymes to form a new, homology-derived model of human CLN2. The primary basis of the homology model was the structure of a complex of sedolisin with a covalently-bound inhibitor, pseudo-iodotyrostatin. Although it has not been directly shown that this compound can inhibit human CLN2, other similar peptides with an aldehyde functionality on their "C-termini" are weak, but detectable, inhibitors of this enzyme (Oda, unpublished). It is thus likely that pseudo-iodotyrostatin or a similar inhibitor might work for CLN2 as well, although the actual contacts between the inhibitor and the enzyme that are seen in the model have to be treated with caution. Another reason for the modeling of a pseudo-iodotyrostatin complex is that CLN2 is a tripeptidase, and that this inhibitor it represents the only experimental structure of a tripeptide analog bound to sedolisin.

The r.m.s. deviation between the corresponding Cα coordinates of the model of CLN2 (Figure 5) and the experimental structure of sedolisin is about 1.75 Å, not much larger than the experimental difference between sedolisin and kumamolisin. Interestingly, the Cys327 and Cys342 residues in the model were found to be ideally positioned to form a disulfide bond even though this was not part of the design strategy. That this bond likely occurs in the real protein is suggested by the fact that these two cysteines are strictly conserved in all known animal species of CLN2 (Figures 1 and 3), although they are absent in all known sequences of bacterial sedolisins. Thus, if this disulfide were experimentally found to exist in CLN2 it would provide support for the correctness of the model.

Comparisons of the substrate binding pockets of CLN2 and sedolisin

Since the principal known activity of CLN2 is that of a tripeptidase, it is expected that three substrate-binding pockets, S1 through S3 (using the nomenclature of Schechter and Berger [16]), should be discernible. Residues P1-P3 of the inhibitor that should occupy these pockets are shown in Figure 6. All the available structures of the complexes of either sedolisin or kumamolisin with inhibitors contain either a tyrosine or a phenylalanine occupying the S1 pocket. Parts of this pocket are fully



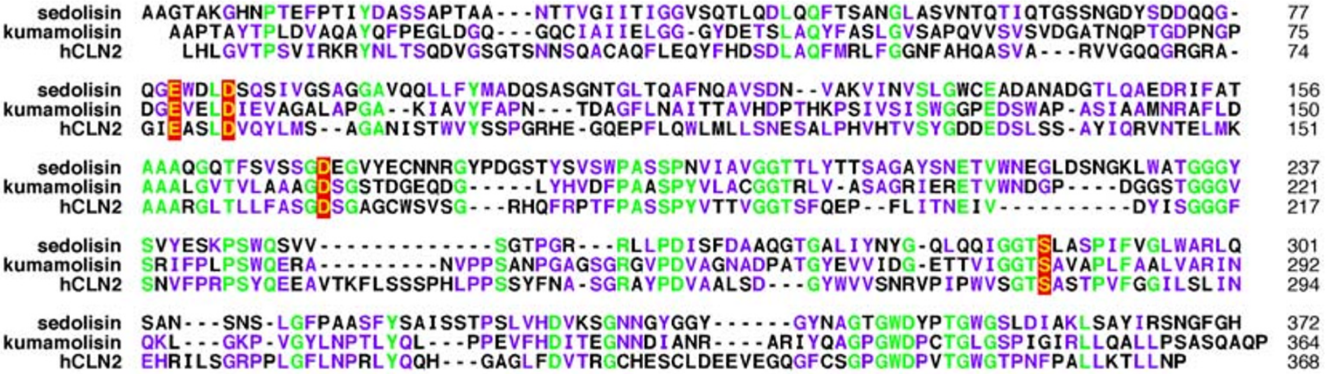
```
Human     LHLGVTPSVIRKRYNLTSQDVGSGTSNNSQACAQFLEQYFHDSDLAQFMRLFGGNFAHQA
Xenopus   ............................................................
Fugu      IHLGVTPSVLRARYNLTSADVGQAQ-NNSQAVAQFLEQYYSPADLVEFMSLFGSGFKHHT
Zebrafish YHLGVTPAVIRNRYNLTAKDVGTAA-NNSQAVAQFLEQYYHPADLAEFMSLFGGGFTHMS

Human     SVARVVGQQGRGRAGIEASLDVQYLMSAGANISTWVYSSPGRHEGQEPFLQWLMLLSNES
Xenopus   ............................................................
Fugu      QVDRVVGTQGAGKAGIEASLDVQYIMSTGANIPTWLFTNPGRHESQEPFLQWMVLLSNMS
Zebrafish TVERVVGTQGGGKAGIEASLDVEYIMSSGANISTWVFTNPGRHESQEPFLQWMLLLSNMS

Human     ALPHVHTVSYGDDEDSLSSAYIQRVNTELMKAAARGLTLLFASGDSGAGCWSVSG-RHQF
Xenopus   ..............DSLSIAFMQRINVEFMKAASRGLTILFASGDDGAGCRHVDKGINTF
Fugu      DLPWVHTISYGDDEDSLSTAYMERINAEFMKAGMRGISLLFASGDSGAGCRRSVK--NSF
Zebrafish AVPWVHTISYGDDEDSLSEAYMNRINIEFMKAGLRGISMLFASGDSGAGCRHLTK--ERN

Human     RPTFPASSPYVTTVGGTSFQEPFLITNEIVDYISGGGFSNVFPRPSYQEEAVTKFLSSSP
Xenopus   RPSFPASSPYVTTVGGTSFKNPFQVTREVSDYISGGGFSNVFPMPDYQAPAVSRYLQSTA
Fugu      RPSFPASSPYVTTVGGTSFKNPFTVSYEVTDYISGGGFSNVFKMPDYQASAVDAYLKTMA
Zebrafish TPSFAASSPYVPTVGGTSFKTPSTLTYEVQILITGGGLNYVFPMPDYQVDAVRAYLKS-V

Human     H-LPPSSYFNASGRAYPDVAALSDGYWVVSNRVPIPWVSGTSASTPVFGGILSLINEHRI
Xenopus   K-LPPQSYYNRSGRAYPDVAALSDNYWVVTNRVPIPWVSGTSASTPVFGGVLSLINDRRL
Fugu      KVLPPQSYYNTTGRAYPDMAALSDNYWVVSNKVPVPWVSGTSASTPVVGGMLSLINDQRL
Zebrafish QSLPPQTYFNTTGRAYPDLAALSDNYWVVSNRVPIPWVSGTSASTPVVGGILSLINDQRF

Human     LSGRPPLGFLNPRLYQQHG---AGLFDVTRGCHESCLDEEVEGQGFCSGPGWDPVTGWGT
Xenopus   HKGLSPLGFLNPALYRLQLNGSQALYDVTEGCHLSCIDDLVL..................
Fugu      LKGLPVLGFLNPRLYRLKG---QALFDVTEGCHLSCLDEQVQGKGFCAAPSWDPVTGWGT
Zebrafish LKGLPALGFINPRLYKMQG---KKLYDVTVGCHLSCLDDKVEGKGFCASPSWDPVTGWGT

Human     PNFPALLKTLLNP
Xenopus   .............
Fugu      PNYPELLTALL--
Zebrafish PNYPVFLASLMD-
```

Figure 3

Sequences of the catalytic domains of CLN2. Complete sequences are shown for CLN2 from human, fugu, and zebrafish, together with the partial sequence of putative CLN2 in Xenopus tropicalis. Residues identical in all four enzymes are colored green and those similar are colored magenta. Active site residues are marked as in Figure 1.

conserved among different sedolisins, whereas other parts of it differ. The right-hand side of the pocket (in the view used in Figure 6) is made of the main chain including residues 164–165 (unless otherwise indicated, the numbering refers to the sequence of the mature human CLN2). This part of the main chain is held in place through a hydrogen-bonded interaction with the side chain of Thr279, part of the signature sequence SGTSAS surrounding the catalytic Ser280 and its equivalents in the other enzymes. Asp165 itself is also conserved since that residue provides the lining of the oxyanion hole, so it can be safely assumed that this part of the S1 pocket is virtually identical. No side chains point into the pocket from there, though, so its importance is limited to providing a steric barrier and excluding solvent. Another wall of the pocket is made up of the main chain of residues 130–132 that flank the conserved Gly131. Again, this part of the chain does not provide any specific interactions with the P1 residue of the substrate.

Considerable differences are seen, however, at the bottom of the pocket, where the side chains of Asp179 in kumamolisin and the equivalent Ser190 in sedolisin make hydrogen bonds to the P1 tyrosine of the inhibitor (if present). The equivalent residue in CLN2 is Thr182, but it is very unlikely that it can assume an orientation that would allow it to make a hydrogen bond to a P1 tyrosine. Other polar residues in the vicinity are Glu175 in sedolisin and the corresponding Asp169 in kumamolisin. However, the residue found here in CLN2 is Cys170, much less



```
   sedolisin  AAGTAKGHNPTEFPTIYDASSAPTAA---NTTVGIITIGGVSQTLQDLQQFTSANGLASVNTQTIQTGSSNGDYSDDQQG-    77
kumamolisin     AAPTAYTPLDVAQAYQFPEGLDGQ---GQCIAIIELGG-GYDETSLAQYFASLGVSAPQVVSVSVDGATNQPTGDPNGP    75
       hCLN2      LHLGVTPSVIRKRYNLTSQDVGSGTSNNSQACAQFLEQYFHDSDLAQFMRLFGGNFAHQASVA---RVVGQQGRGRA-    74

   sedolisin  QGEWDLDSQSIVGSAGGAVQQLLFYMADQSASGNTGLTQAFNQAVSDN--VAKVINVSLGWCEADANADGTLQAEDRIFAT   156
kumamolisin  DGEVELDIEVAGALAPGA--KIAVYFAPN---TDAGFLNAITTAVHDPTHKPSIVSISWGGPEDSWAP-ASIAAMNRAFLD   150
       hCLN2  GIEASLDVQYLMS--AGANISTWVYSSPGRHE-GQEPFLQWLMLLSNESALPHVHTVSYGDDEDSLSS-AYIQRVNTELMK   151

   sedolisin  AAAQGQTFSVSSGDEGVYECNNRGYPDGSTYSVSWPASSPNVIAVGGTTLYTTSAGAYSNETVWNEGLDSNGKLWATGGGY   237
kumamolisin  AAALGVTVLAAAGDSGSTDGEQDG-----LYHVDFPAASPYVLACGGTRLV-ASAGRIERETVWNDGP----DGGSTGGGV   221
       hCLN2  AAARGLTLLFASGDSGAGCWSVSG---RHQFRPTFPASSPYVTTVGGTSFQEP--FLITNEIV----------DYISGGGF   217

   sedolisin  SVYESKPSWQSVV-------------SGTPGR---RLLPDISFDAAQGTGALIYNYG-QLQQIGGTSLASPIFVGLWARLQ   301
kumamolisin  SRIFPLPSWQERA----------NVPPSANPGAGSGRGVPDVAGNADPATGYEVVIDG-ETTVIGGTSAVAPLFAALVARIN   292
       hCLN2  SNVFPRPSYQEEAVTKFLSSSPHLPPSSYFNA-SGRAYPDVAALSD---GYWVVSNRVPIPWVSGTSASTPVFGGILSLIN   294

   sedolisin  SAN---SNS-LGFPAASFYSAISSTPSLVHDVKSGNNGYGGY------GYNAGTGWDYPTGWGSLDIAKLSAYIRSNGFGH   372
kumamolisin  QKL---GKP-VGYLNPTLYQL---PPEVFHDITEGNNDIANR----ARIYQAGPGWDPCTGLGSPIGIRLLQALLPSASQAQP 364
       hCLN2  EHRILSGRPPLGFLNPRLYQQH---GAGLFDVTRGCHESCLDEEVEGQGFCSGPGWDPVTGWGTPNFPALLKTLLNP   368
```

Figure 4

Sequence alignment of bacterial and mammalian enzymes. Alignment of the sequences of sedolisin, kumamolisin, and human CLN2 used in the construction of the model of the latter enzyme. The colors scheme is the same as in Figure 2.

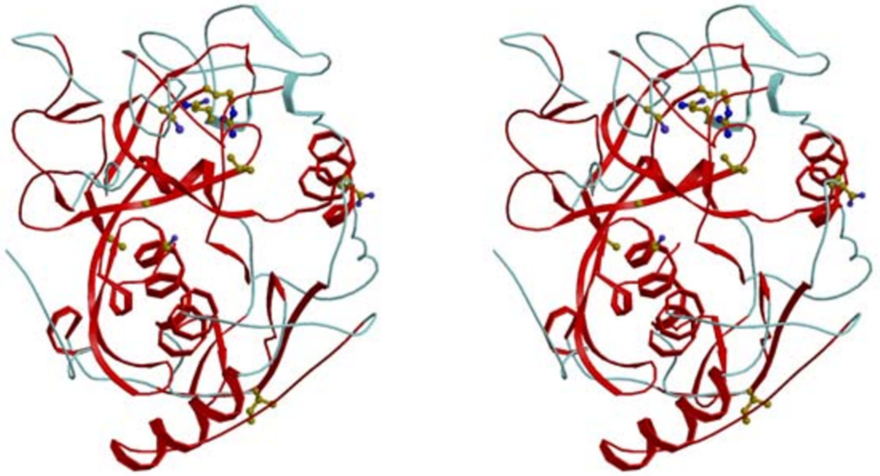


Figure 5

A homology-derived model of human CLN2. Ribbon diagram of the Cα trace of CLN2, with the segments that were modeled based on the highly conserved core of sedolisin and kumamolisin (r.m.s. deviation of 1 Å) colored in red. Side chains of the residues that were found to be mutated in the genes of families of patients with late-infantile neuronal ceroid lipofuscinosis [17] are marked in ball-and-stick.

polar than its counterparts. There is also no equivalent to the polar interaction between Glu175 and Glu171 in sedolisin, since the equivalent of the latter residue in CLN2 is Ser166, much smaller and pointing away.

The side of the S1 pocket that is created by the very flexible side chain of Arg179 in sedolisin contains only a much smaller Ser174 in CLN2, and thus is much more open in the latter protein. This part of the pocket, with the main chain of the protein quite distant from the substrate, is indeed not well conserved in these proteins, with kumamolisin missing it entirely due to a deletion in the corresponding sequence position. In summary, the S1 pocket in CLN2 has less polar character than the equivalent pocket in the related proteins, and is lacking direct polar anchors for any side chains that might be present in the substrate.

The S2 pocket in CLN2 is also quite open and accessible to solvent. It is most likely larger than the equivalent

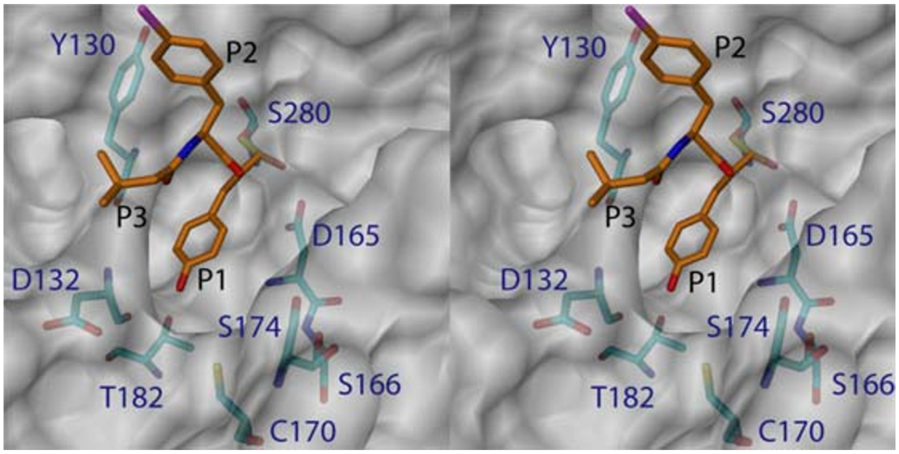


Figure 6
A model of the active site of human CLN2. The enzyme is shown in complex with pseudo-iodotyrostatin, a good inhibitor of the sedolisin family of peptidases. Only selected residues of the enzyme are explicitly shown on the background consisting of the molecular surface. The stick model of the inhibitor is colored gold and the P1-P3 residues are labeled in black. Similar views have been previously published for the experimentally-determined structures of sedolisin and kumamolisin [8]. The figure was prepared using the program DINO http://www.bioz.unibas.ch/~xray/dino.

pockets in either sedolisin or kumamolisin, since these are limited by Trp81 in the former and Trp129 in the latter (these residues originate from different parts of the backbone in the two enzymes and are not topologically related). Tyr130, an equivalent of the latter residue in CLN2, is unlikely to come into direct contact with the P2 residue of the substrate due to its greater distance (almost 4 Å for the closest atoms).

An important remaining puzzle is that the predicted structure of CLN2 does not show any clear limitations of the S3 pocket that could explain the tripeptidase activity of this enzyme. The location of the P3 side chain of the substrate is ambiguous, since it could point in either one of two directions by exchange with the N-terminal amine. The only negatively charged residue of CLN2 that is found in this vicinity is Asp132. Although in the current model the distance between its carboxylate and the nitrogen of the N-terminus of the modeled substrate is about 6 Å, these two groups could be brought into hydrogen-bonding range by some allowed changes in the torsion angles of the protein. Such a conformational change would involve breaking of the hydrogen bond between Asp132 and Ser139. However, this latter interaction is not likely to be structurally crucial since the serine is not absolutely conserved in all CLN2-like enzymes.

Location of mutations implicated in disease

Location of mutations found through a genetic survey of families of classic late-infantile neuronal ceroid lipofuscinosis patients has been described previously [17]. Most such mutations result in expression of either truncated enzyme or in incorrect intron-exon splices. However, some of the mutations lead to single amino acid substitutions in the mature enzyme. Such mutations include I92N, E148K, C170R and C170Y, V190D, G194E, Q227H, R252H, A259E, and S280L (Figure 5). Only the role of the latter mutation is completely clear, since it replaces the catalytic serine of CLN2 with a side chain that cannot support its enzymatic activity. No other residues appear to be located in the immediate vicinity of the substrate. Residues Val190, Gly194, and Arg252 are very highly conserved not only in CLN2 but also in other sedolisins and must play an important structural role. The reasons why the remaining mutations would lead to the loss of enzymatic activity are much less clear, but the wide distribution of these mutations in the structure supports the conclusion that any modifications to CLN2 that would abolish or impair its function could lead to the development of the disease.

Substrates and inhibitors of CLN2

Little is known at this time about biologically-relevant substrates of CLN2. Various defects that include truncations and single-site mutations in CLN2 have been found in the genes of patients that display symptoms of late-infantile neuronal ceroid lipofuscinosis [17]. One of the symptoms of the disease is the accumulation of an autofluorescent material, ceroid-lipofuscin, in lysosomal storage bodies in various cell types, primarily in the nervous system. Since a major component of such bodies appears to be intact subunit *c* of mitochondrial ATP synthase, this protein has been implicated as a potential biological target of the protease. It has been shown recently that CLN2 can indeed degrade this subunit on its N terminus [18], but the unambiguous proof that this is indeed the most important target is still lacking. CLN2 is capable of processing a number of different angiotensin-derived peptides [19], with the efficiency of cleavage dependent on the length of such peptides. The most efficiently processed peptide consisted of 14 amino acids, with the tripeptide Asp-Arg-Val removed from its N terminus. The model of CLN2 presented here can easily accommodate this peptide on the P side of the substrate-binding site, although the exact mode of binding of the long P' portion of the substrate remains obscure. The observation that an analogous peptide acetylated on its N terminus cannot be processed supports the postulate that the interactions of the N-terminal amino group with the side chain of Asp132 may be the most important feature defining the tripeptidase specificity of CLN2. A number of different tripeptides can be serially processed from glucagon, with their sequences varying widely [20]. Again, however, all of these tripeptides can be easily accommodated in the substrate-binding site of the CLN2 model. Other potentially biologically relevant substrates include cholecystokinin and possibly other neuropeptides [21].

An intriguing property of CLN2 is its reported ability to cleave collagen-related peptides [22]. The tripeptides resulting from such processing include Gly-Pro-Met, Gly-Pro-Arg, and Gly-Pro-Ala. It has been recently reported that kumamolisin, and particularly a closely related protein from *Alicyclobacillus sendaiensis* (kumamolisin-*As*) can efficiently cleave not only collagen-related peptides, but also native type I collagen [23]. With the substrate-binding site of CLN2 resembling that of kumamolisin more than sedolisin (the latter enzyme has low, if any, collagenase activity), the potential collagen-processing role of CLN2 might warrant further investigation.

Conclusions

Since the catalytic machinery of CLN2 matches closely that of sedolisin, kumamolisin, and other members of the family of serine-carboxyl peptidases, the enzymatic mechanism of all these enzymes is most likely the same. Design of inhibitors specific for CLN2 should incorporate the features that have been proven to be important for the related enzymes, such as the placement of an aldehyde functionality capable of making covalent interactions with the catalytic serine, or the utilization of chloromethyl ketone for the same purpose. Since the few inhibitors that have been successfully used in the studies of sedolisins are either longer than tripeptides or contain blocking groups on their N termini, new tripeptide-based inhibitors with free N termini are now being synthesized (Oda, unpublished). It will be necessary to test the binding properties of different substrates in order to determine the most promising peptide sequences. Analysis of the model of CLN2 suggests that the size of the S1 subsite is much larger than in either sedolisin or kumamolisin, and thus the use of a large P1 group might be indicated. Of course, the availability of an experimental crystal structure will make the design of inhibitors easier and we are continuing our efforts to crystallize CLN2 from different sources.

Methods

Homology modeling

Three-dimensional, atomic-scale models of CLN2 were developed by exploiting the sequence similarity to the sedolisin and kumamolisin proteins (r.m.s. deviation of 1.0 Å for 273 pairs of Cα atoms in the core of the enzymes). Presently, these two enzymes are the only members of the newly-defined sedolisin/serine-carboxyl peptidase family [8] for which the crystal structures have been published [5–7]. The actual Protein Data Bank [24]: http://www.pdb.org/) entries used in the modeling were 1GA4.pdb and 1GT9.pdb for sedolisin and kumamolisin, respectively.

The first step was to form a global, multiple sequence alignment between all known members of the sedolisin family. Studies have shown that incorporating the specific patterns of amino acid residue-type variation and conservation among a family of homologous proteins provides superior results over simple, pair-wise sequence alignment [25]. Sequence files representing the different subfamilies were extracted from the non-redundant GenBank database [26] using sedolisin, kumamolisin, and the human CLN2 sequences as queries to the web-based version of the BLAST program [27]: http://www.ncbi.nlm.nih.gov/BLAST/. Initial multiple sequence alignments were formed with the ClustalX computer program [28]. As is expected for a family of proteins, highly-conserved segments were found aligned to the crystal structure-identified core regions of the sedolisin and kumamolisin sequences. Subsequently, the sequences were divided into two groups: those closer to sedolisin than kumamolisin and *vise-a-versa*. The alignment of these two groups was then manually set by the observed structural alignment of the sedolisin and kumamolisin proteins. Finally, some additional adjustment was required to correct the few places where highly conserved residues of the core regions were slightly out of alignment among different subfamilies of sequences.

The model of human CLN2 was built using the structure of sedolisin complexed with the inhibitor pseudo-tyrostatin [5,6] as a template. The reason for this choice is that while different protein models were generally comparable, the chosen inhibitor was most compatible with the tripeptidase character of CLN2. With the correspondence of residues specified in the alignments, atomic coordinates were transferred to the target sequence by a variety of methods, including the homology modeling modules of the Look/GeneMine [29] and DeepView [30] computer program packages. For the core and active site of the protein, coordinates for identical residues were simply transferred unchanged; whereas, special care was required to position the side chains of residues differing from the template. This was first accomplished automatically by the two computer packages, then manually adjusted in the Quanta molecular modeling package (Accelrys, Inc.) to better mimic the templates and optimize the interactions with surrounding residues. A similar two-step approach was used to manifest the insertions and deletions in the variable, loop regions of the protein, where it was necessary to create new backbone as well as side chain coordinates for the models. It should be noted that, for obvious reasons, the conformation of poorly conserved loop regions is generally the least accurate aspect of a homology model. Fortunately, these problematic loops will not significantly affect the active site of the model, since only two of them impinge on the boundary of this highly conserved, functional region.

Refinement and analysis of the model

The model was finished by performing energy minimization *in vacuo* with the computer program CHARMM [31]. This refined the structure by bringing the covalent geometry and non-bonded interactions into agreement with experimentally observed and calculated values. Such optimizations included adjusting bond lengths, 3-point angles and 4-point dihedral angles, as well as eliminating atomic overlap and forming salt-bridges and hydrogen bonds. Since presently the potential energy functions used to describe the atomic-scale models are not sufficiently comprehensive and accurate, the final energy of the model was not used as an indicator of the realistic quality of the structure. The final quality of the structure was analyzed with the computer program PROCHECK [32]. The structure was deposited at the PDB under accession code 1R60.

Authors' contributions

AW initiated this project and analyzed the genomic distribution of this family of enzymes. SRD contributed the modeling of the three-dimensional structure of CLN2. ML analyzed the model and compared it to the crystal structures of sedolisins. HO, KO and BMD contributed their experience gained from studies of serine-carboxyl peptidases and the design of their inhibitors, aimed at analysis of substrate-enzyme interactions and enzyme specificity. All authors read and approved the final manuscript.

Acknowledgements

Extensive discussions with Dr. A. Barrett (Wellcome Trust Sanger Institute, Hinxton, UK) are gratefully acknowledged. Modeling efforts were facilitated by Dr. Robert Jernigan (Iowa State University). This work was supported in part by a Grant-in-Aid for Scientific Research (B), 15380072, from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to K.O.); by NIH grants DK18865 and AI39211 (to B.M.D.); and in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract No. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products or organizations imply endorsement by the U. S. Government.

References

1. Ellis S: **Studies on the serial extraction of pituitary proteins.** *Endocrinology* 1961, **69:**554-570.
2. Sleat DE, Donnelly RJ, Lackland H, Liu CG, Sohar I, Pullarkat RK and Lobel P: **Association of mutations in a lysosomal protein with classical late- infantile neuronal ceroid lipofuscinosis.** *Science* 1997, **277:**1802-1805.
3. Rawlings ND and Barrett AJ: **Tripeptidyl-peptidase I is apparently the CLN2 protein absent in classical late-infantile neuronal ceroid lipofuscinosis.** *Biochim Biophys Acta* 1999, **1429:**496-500.
4. Lin L,, Sohar I,, Lackland H, and Lobel P: **The human CLN2 protein/tripeptidyl-peptidase I is a serine protease that autoactivates at acidic pH.** *J Biol Chem* 2001, **276:**2249-2255.
5. Wlodawer A, Li M, Dauter Z, Gustchina A, Uchida K, Oyama H, Dunn BM and Oda K: **Carboxyl proteinase from Pseudomonas defines a novel family of subtilisin-like enzymes.** *Nature Struct Biol* 2001, **8:**442-446.
6. Wlodawer A, Li M, Gustchina A, Dauter Z, Uchida K, Oyama H, Goldfarb NE, Dunn BM and Oda K: **Inhibitor complexes of the Pseudomonas serine-carboxyl proteinase.** *Biochemistry* 2001, **40:**15602-15611.
7. Comellas-Bigler M, Fuentes-Prior P, Maskos K, Huber R, Oyama H, Uchida K, Dunn BM, Oda K and Bode W: **The 1.4 Å crystal structure of kumamolysin: a thermostable serine- carboxyl-type proteinase.** *Structure* 2002, **10:**865-876.
8. Wlodawer A,, Li M,, Gustchina A,, Oyama H,, Dunn BM, and Oda K: **Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases.** *Acta Biochim Polon* 2003, **50:**81-102.
9. Lee BR, Furukawa M, Yamashita K, Kanasugi Y, Kawabata C, Hirano K, Ando K and Ichishima E: **Aorsin, a novel serine proteinase with trypsin-like specificity at acidic pH.** *Biochem J* 2003, **371:**541-548.
10. Nishii W, Ueki T, Miyashita R, Kojima M, Kim YT, Sasaki N, Murakami-Murofushi K and Takahashi K: **Structural and enzymatic characterization of physarolisin (formerly physaropepsin) proves that it is a unique serine-carboxyl proteinase.** *Biochem Biophys Res Commun* 2003, **301:**1023-1029.
11. Weber IT, Miller M, Jaskólski M, Leis J, Skalka AM and Wlodawer A: **Molecular modeling of the HIV-1 protease and its substrate binding site.** *Science* 1989, **243:**928-931.
12. Anand K,, Ziebuhr J,, Wadhwani P,, Mesters JR, and Hilgenfeld R: **Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs.** *Science* 2003, **300:**1763-1767.
13. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A,

Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC and Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.

14. Katz ML and Johnson GS: **Mouse gene knockout models for the CLN2 and CLN3 forms of ceroid lipofuscinosis.** *Eur J Paediatr Neurol* 2001, **5 Suppl A:**109-114.
15. Orry A, and Wallace BA: **A proposed model for the late-infantile neuronal ceroid lipofuscinosis (Batten Disease) protein CLN2.** *Protein Pept Lett* 1999, **6:**1-5.
16. Schechter I and Berger A: **On the size of the active site in proteases. I. Papain.** *Biochem Biophys Res Commun* 1967, **27:**157-162.
17. Sleat DE, Gin RM, Sohar I, Wisniewski K, Sklower-Brooks S, Pullarkat RK, Palmer DN, Lerner TJ, Boustany RM, Uldall P, Siakotos AN, Donnelly RJ and Lobel P: **Mutational analysis of the defective protease in classic late-infantile neuronal ceroid lipofuscinosis, a neurodegenerative lysosomal storage disorder.** *Am J Hum Genet* 1999, **64:**1511-1523.
18. Ezaki J, Takeda-Ezaki M and Kominami E: **Tripeptidyl peptidase I, the late infantile neuronal ceroid lipofuscinosis gene product, initiates the lysosomal degradation of subunit c of ATP synthase.** *J Biochem (Tokyo)* 2000, **128:**509-516.
19. Warburton MJ and Bernardini F: **The specificity of lysosomal tripeptidyl peptidase-I determined by its action on angiotensin-II analogues.** *FEBS Lett* 2001, **500:**145-148.
20. Vines D and Warburton MJ: **Purification and characterisation of a tripeptidyl aminopeptidase I from rat spleen.** *Biochim Biophys Acta* 1998, **1384:**233-242.
21. Bernardini F, and Warburton MJ: **Lysosomal degradation of cholecystokinin-(29-33)-amide in mouse brain is dependent on tripeptidyl peptidase-1: implications for the degradation and storage of peptides in classical late-infantile neuronal ceroid lipofuscinosis.** *Biochem J* 2002, **366:**521-529.
22. McDonald JK, Hoisington AR and Eisenhauer DA: **Partial purification and characterization of an ovarian tripeptidyl peptidase: a lysosomal exopeptidase that sequentially releases collagen-related (Gly-Pro-X) triplets.** *Biochem Biophys Res Commun* 1985, **126:**63-71.
23. Tsuruoka N, Nakayama T, Ashida M, Hemmi H, Nakao M, Minakata H, Oyama H, Oda K and Nishino T: **Collagenolytic serine-carboxyl proteinase from Alicyclobacillus sendaiensis strain NTAP-1: Purification, characterization, gene cloning, and heterologous expression.** *Appl Environ Microbiol* 2003, **69:**162-169.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
25. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T and Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284:**1201-1210.
26. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA and Wheeler DL: **GenBank.** *Nucleic Acids Res* 2002, **30:**17-20.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
28. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
29. Lee C, and Irizarry K: **The GeneMine System for genome/proteome annotation and collaborative data mining.** *IBM Systems Journal* 2001, **40:**592-603.
30. Guex N and Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
31. Brooks BR,, Bruccoleri RE,, Olafson BD,, States DJ,, Swaminathan S, and Karplus M: **CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.** *J Comp Chem* 1983, **4:**187-217.
32. Laskowski RA, MacArthur MW, Moss DS and Thornton JM: **PROCHECK: program to check the stereochemical quality of protein structures.** *J Appl Crystallogr* 1993, **26:**283-291.