amplification was performed using 5 µL of the first reaction and Oligo dC14-Sal I and Rev. 4 primers under the same conditions as in 3' RACE.

All PCR products were ligated into pGEM-T vector and transformed into *Escherichia coli* DH5-α according the supplier's instructions. Mini-preparations of the cloned products were prepared using the Wizard 373 plasmid preparation kit according to the manufacturer's specifications. The DNA was digested with Pst I and electrophoresed in agarose according to standard methods to verify inserts [22]. The inserts were sequenced by automated sequencing using M13 forward and reverse primers and internal primers derived from the cloned sequences on an ABI 373A DNA sequencer.

The full-length cDNA was amplified using the 5'-term and 3'-term primers (Table 1). The product was gel-purified using the Geneclean III procedure as above and sequenced directly using the internal primers and the same automated sequencing procedure (Figure 2).

Sequence Analysis:

Open reading frames for all sequences were identified using the Baylor College of Medicine (BCM) Search launcher Sequence Analysis tools. The final protein sequence was aligned with related enzyme amino acid sequences using the Feng and Doolittle Propack programs[20] with the Dayhoff 250 and Blosum 62 amino acid similarity matrices. Structural comparisons with the cyanogenic β-glucosidase from white clover (1CBG)[15] were made using the SwissProt PDB viewer[23] to thread the sequence of the *D. cochinchinensis* β-glucosidase into this structure and resolve atomic overlaps. The resultant structure was subjected to energy minimization at the Swiss Model server using the default parameters. Further refinement of

dalcochinase catalytic-site side chains was done by manually adjusting torsion angles to match the 1CBG structure, while maintaining low energy conformations.

RESULTS:

Protein sequencing and analysis

The dalcochinase enzyme was purified from seeds of *Dalbergia cochinchinensis* Pierre by previously described methods which produce protein that is homogeneous by SDS-PAGE and native gel electrophoresis [11]. The protein was deglycosylated with Endo-F/Endoglycosidase F, which increased its electrophoretic mobility on SDS-PAGE. After purification by HPLC on a C8 column, the deglycosylated protein was sequenced by automated Edman degradation to yield the N-terminal sequence: IDFAKEVA. Fractions of the deglycosylated protein were also digested with Endopeptidase Lysine-C and Trypsin, and the resultant peptides were separated by HPLC on a C8 column as shown in Figure 1 A and B, respectively. The peak fractions were collected and several were subjected to automated Edman degradation. The sequences determined are indicated in Figure 1 beneath the peptide maps.

Blast comparison of peptide sequences to the NCBI nr sequence database, indicated several homologies with known sequences of β-glucosidases and related proteins. In particular, the Endo Lys-C peak 9 peptide was identical at 12 out of 13 positions with the sequence of a thioglucosidase (myrosinase 3D) from *Sinapsis alba* which contains the catalytic nucleophile. This analysis indicated that the dalcochinase enzyme was likely to be a β-glucosidase from glycosyl hydrolase family 1 [2]. Several sequences from this family were retrieved from the on-line databases and

aligned and the peptide sequences were positioned in the multiple alignment to determine the approximate distance between peptides. However, the amino terminal was not significantly homologous to other sequences, so it could not be positioned from the alignment. Degenerate oligonucleotide primers were designed by back-translating the peptide sequence of the amino terminal to give the For.1 primer, the TP7 and Lys9 peptides (in the antisense direction) to give the Rev.1 and Rev.2 primers, respectively (Table I, Figure 2).

To determine the best tissue to extract RNA for PCR cloning, crude enzyme was extracted from several tissues and the relative amounts of β-glucosidase and β-fucosidase activities were determined. Both dried seeds, from which the enzyme was originally extracted, and germinated seeds had greater than 100 times the level of β-glucosidase activity found in leaf, stem, and shoot tip and approx. 450 times the levels of β-fucosidase activity found in these tissues. They each had a 2.2 fold β-fucosidase activity compared to β-glucosidase activity, in agreement with the previously characterized enzyme purified from dried seeds [10,11]. Leaves, stems, and shoots, in contrast, had much lower levels of β-fucosidase relative to β-glucosidase, indicating that other β-glucosidases may predominate in these tissues. Though germinated seed and dry seed had similar amounts of enzyme per mass, the germinated seed had twice the activity per seed, indicating new enzyme was synthesized or activated during germination. Therefore, the germinated seeds were used to prepare RNA for RT-PCR cloning of the β-glucosidase cDNA.

RNA prepared from germinated *Dalberia cochinchinensis* seeds was reverse-transcribed from the Rev.2 primer, and the resultant single-strand cDNA was amplified with the For.1 and Rev.1 primers. A specific product was identified by

PAGE, eluted from the gel, cloned into a pGEM-T plasmid, and sequenced. The insert sequence contained 224 bp, including the primer sequences, which translated to 74 amino acids at the amino terminal of the mature β-glucosidase sequence. The fragment is indicated schematically in Figure 2. The protein sequence included the Lys4 and TP23 peptide sequences in addition to the N-terminal and TP7 sequences used to design the For.1 and Rev.1 primers.

3′ and 5′ RACE

In order to determine the sequence of the 3′ end of the cDNA, 3′ rapid amplification of cDNA ends (RACE) was conducted starting with reverse-transcription of RNA from germinated *D. cochinchinensis* seeds with the poly-T primer, AP (shown in Table 1). Hemi-nested PCR was then conducted with the For. 2 and For. 3 primers, derived from the original cDNA fragment sequence and the AUAP primer, contained in the AP sequence as shown in Figure 2. The first amplification produced a diffuse band centered at approx. 2000 bp on agarose gel electrophoresis, while the hemi-nested PCR produced a slightly smaller product. The second product was gel-purified, cloned, and sequenced (Figures 2 & 3). The clone contained 1,735 bp, including an uninterrupted reading frame of 473 amino acids, starting from the For. 3 primer at the 5′ end and ending with a TGA stop codon. After the stop codon, there were 292 nucleotides of putative 3′ untranslated region before the polyA tail, corresponding to the reverse transcription primer. The sequence included the 3′ end of the initial PCR fragment, except that some of the nucleotides in the area of the Rev.1 primer used to amplify that fragment did not match the new sequence. Notably, the amino acid sequence had changed to VDQFHRYK, instead of the VDQFHYYN of the TP7 peptide sequence used to design the primer. The

sequence also included the sequences of the Lys3, Lys4, Lys5, Lys6, Lys9, Lys10, TP9, TP22, and TP37 peptides (see Figure 3).

In order to determine the sequence of the amino terminal of the protein, 5' RACE was employed. Total RNA was reverse-transcribed using the Rev.3 primer, designed from the sequence of the initial PCR product (Table 1, Figure 2). The single strand cDNA was tailed with dGTP, and hemi-nested PCR was performed with the oligo-dC_{17}-SalI primer and the Rev. 3 and Rev. 4 primers (see Figure 2). The second reaction produced a 210 bp 5' cDNA fragment, which was cloned and sequenced and found to contain 193 base pairs 3' of the poly C tract of the Oligo-dC-SalI primer. When translated, the protein sequence contained 75 amino acids, including the N-terminus of the mature protein and a 23 amino acid pre-sequence, starting from the first methionine (Figure 3). The pre-peptide included a stretch of hydrophobic amino acids, typical of a signal sequence for transport into the endoplasmic reticulum.

Direct Sequencing of the PCR-Amplified cDNA

In order to ensure that the cloned sequences were correct, a full coding sequence cDNA was amplified with the 5'-term and 3'-term primers, designed from the 5' and 3' RACE product clones. This 1906 bp PCR product was directly sequenced in both directions at least twice from independent PCR amplifications. The complete sequence, including the cloned sequences from outside the 5'term and 3' term primers is shown in Figure 3 (Genbank accession AF163097).

Analysis of Peptide Sequence

As noted, the TP7 peptide sequence did not match the DNA sequence for this segment at its carboxyl-terminal, except in the initial clone, where this region was provided by the Rev.1 primer designed from the TP7 peptide sequence. Inspection of

the translated cDNA sequence indicated another predicted peptide, starting from the valine at residue 127 with the sequence YYN in positions 13 to 15 (as in TP7). Closer inspection of the protein sequencing chromatographs indicated minor peaks in cycles 1-12 consistent with this sequence: VSGGINQTGVDYYN (peptide T7B in Figures 1 and 3) and a small arginine peak at cycle 13 (consistent with the TP7A sequence of Figures 1 and 3). However, no asparagine peak was seen in the sixth cycle. MALDI-TOF mass spectroscopy of the HPLC peak fraction indicated two peaks at 1345.5 amu and 1846.9 amu, consistent with the sequences SNGDVAVDQFHR (TP7A) and VSGGINQTGVDYYN (TP7B) plus one N-acetyl-glucosamine (or N-acetyl-galactosamine), respectively. No other major peaks were observed in the mass spectrum of this fraction.

Sequence Analysis.

The protein sequence translated from the cDNA sequence was aligned to the known sequences of glycosyl hydrolase family 1, which contains the related plant β-glucosidases[2,14]. An alignment of some selected members of this family is shown in Figure 4. The mature protein showed highest percent identities with the cyanogenic and noncyanogenic β-glucosidases from white clover (60% and 56%, respectively), and the black cherry amygdalin hydrolase isoform AH I (56%). The alignment showed that the glutamic acids which have previously been identified as the catalytic nucleophile and catalytic acid are conserved in the *D. cochinchinensis* protein at positions E 396 and E 182, respectively of the mature protein (E 419 and E 205 of the precursor protein). Other residues reported to be near the active site of the white clover cyanogenic β-glucosidase were largely conserved in the dalcochinase. Of the residues shown by Barret *et al.*[15] to surround the active site, only Val 254 of

the white clover enzyme is replaced (by His) in the dalcochinase. A model of the active site based on the white clover protein shows that this residue is close to the catalytic proton donor, where Val is thought to provide a hydrophobic pocket in the white clover enzyme (Figure 6). The alignment and model also showed that all the cysteines found in the 1CBG structure are conserved and additional cysteines are found at positions 278, 338 and 339 in the mature dalcochinase.

A phylogenetic tree of the Family 1 glycosyl hydrolases was constructed and is represented in the phenogram in Figure 5. Dalcochinin glucoside β-glucosidase is clustered with the dicotyledon defense-related β-glucosidases, such as linamarase, and amygdalin and prunasin hydrolases. Other plant β-glucosidases and thioglucosidases were next most closely related, followed by the mammalian enzymes, such as lactase phlorizin hydrolase and the cytoplasmic β-glucosidase (not shown).

Discussion:

This paper reports the cDNA and amino acid sequence of the dalcochinin-8'-O-β-glucoside β-glucosidase, dalcochinase, from the Thai Rosewood, *Dalbergia cochinchinensis* Pierre. Possible N-linked glycosylation sites were identified and one site was confirmed by Edman degradation and mass spectroscopy. The amino acid sequence indicates that this protein belongs to the glycosyl hydrolase family 1 and a preliminary structural model is predicted based on the previous x-ray crystallographic model of the closely related cyanogenic β-glucosidase from white clover [15].

The cDNA from which the sequence was determined was isolated by polymerase chain reaction (PCR). Because of the relative imprecision of the Taq polymerase used in this technique, the accuracy of the sequence was ensured by using

direct sequencing of the PCR product to determine the sequence of the coding region. Direct sequencing should eliminate the concern of mistakes made in the PCR which are then found in individual clones, since any mistake should represent a relatively small portion of the strands being sequenced from the entire PCR product [24]. Use of two independent PCR reactions for sequencing provides additional insurance against mistakes made during early cycles of PCR.

Proof that the sequence is correct and corresponds to the dalcochinase is also seen in that the predicted amino acid sequence matched the amino terminus and proteolytic peptides sequenced by Edman degradation. Only one peptide, TP7, did not correctly match the amino acid sequence predicted from the nucleotide sequence at its carboxyl terminus, which ended in YYN, instead of the predicted R (Figures 1 & 3). Secondary inspection of the automated sequencing cycle chromatographs indicated that there were actually two signals, one corresponding to the predicted peptide at the N-terminus: SNGDVAVDQFHR (TP7A), and one corresponding to another predicted peptide with the TP7 C-terminus: VSGGINQTGVDYYN (TP7B), except that the asparagine peak was missing at cycle 6 (Asn109 in the mature protein). As Asn109 is a possible glycosylation site, it might be expected to be missing. Mass-spectroscopy of the HPLC fraction containing the TP7 peak confirmed there were 2 signals corresponding to the peptides TP7A and TP7B with a mass greater than predicted from its amino acid composition. The facts that the mass corresponded to 1 N-acetyl-glucosamine, that the protein was deglycosylated by Endo F, which leaves one glucosamine, and that the Edman degradation did not provide the predicted Asn peak at the 6th cycle, suggests that Asn109, is indeed a site of N-linked glycosylation. It could be noted that N-acetyl glucosamine and N-acetyl-galactosamine have the

same mass and the peptide could be O-glycosylated, except for the release of Thr at cycle 8 (the only possible site of O-linked glycosylation in this peptide). Also, an O-linked glycosylation site would be expected to have more than one sugar residue, since no endoglycosidase specific for O-linked carbohydrates was used. Therefore, the asparagine at position 109 is likely to be glycosylated.

An increase in electrophoretic mobility on SDS-PAGE after treatment with Endo F (from an apparent MW of 66,000 Da to 63,000 Da) also indicated that dalcochinase is a glycoprotein, which passes through the endoplasmic reticulum and is N-glycosylated. Thus, the protein is likely localized in the secretory organelles, vacuole, or outside the cells. A total of 8 possible sites of N-linked glycosylation were identified in the sequence (Figure 3). Modeling of the protein structure based on the x-ray crystallographic structure of the white clover cyanogenic β-glucosidase (PDB code 1CBG [15]), indicated that all possible sites had some level of surface exposure. This number of sites is the same as the 8 sites seen in the white mustard myrosinase crystal structure and significantly more than the 4 possible sites seen in the 1CBG protein [25]. As both dalcochinase and white mustard myrosinase are found in seeds, while 1CBG is found in white clover leaves, this supports the suggestion of Burmeister *et al.* that the high level of glycosylation may support solubility in the relatively dry seed environment.

Multiple sequence alignment and phylogenetic analysis indicated that dalcochinin β-glucosidase belongs to glycosyl hydrolase family 1[2] and is most closely related to β-glucosidases involved in dicotyledon defense mechanisms. Highest percent identities were seen with white clover cyanogenic β-glucosidase (60% identity) and black cherry amygdalin hydrolase (56% identity). Additionally,

the peptide containing the catalytic nucleophile (i.e. the Lys9 peptide) was almost identical to the corresponding peptide from white mustard myrosinase 3D (identical at 12 of 13 residues). However, when the sequence was compared to the SWISSProt Prosite consensus sequences for family 1 enzymes[26,27], they did not match exactly, suggesting that these sequences need to be modified slightly. Residue Glu61 is at a position where Prosite pattern only allows Gly, Ser, Thr, or Ala. Tyr416 and Asp423 are also not allowed residues for their positions according to the Prosite pattern. These are minor changes, since the Prosite consensus sequences are degenerate at these positions, but they indicate the value of having more sequence data available for developing such consensus libraries.

Glutamate 396 was postulated to be the catalytic nucleophile due to its homology to Glu 358 in the β-glucosidase from *Agrobacterium*, which has been shown to act as the catalytic nucleophile.[28] Similarly, Glu 182 was homologous to Glu198 of the cassava cyanogenic β-glucosidase, which was shown to be the catalytic proton donor in that enzyme.[29] Though these exact positions have not been shown experimentally, the carboxylic acid residues are in line with the inhibition of the enzyme by mercuric compounds and CBE.[11,12]

Due to the high sequence identity with the cyanogenic β-glucosidase from white clover, the structure of dalcochinase is expected to be very similar to the crystallographic model for that protein. The sequence was therefore applied to that model and energy minimization was performed to eliminate atomic overlaps using the Swiss Model system [23]. This approach was recently used to identify the catalytic residues of the barley β-glucosidase BII [30]. The model produced had an active site which was superimposable with that of the 1CBG structure (Figure 6). Very few

changes were observed in this active site, and the catalytic nucleophile, Glu395, and catalytic acid base, Glu183, were observed in their expected positions near the carboxyl ends of β-barrel strands 7 & 4, respectively [31], supporting their purported roles. Of the residues identified by Barrett *et al.*[15] as participating in the active site, the only substitution was of a Val with His at His253 of the mature dalcochinase enzyme (Figure 6 c, d). This histidine appears to hydrogen bond to Glu182 in the model. This, and the change of Trp to Ser at Ser184 may result in a less hydrophobic environment for Glu182, which was proposed to be in a hydrophobic pocket in the 1CBG structure. This may be reflected in the optimal pH of 5 for dalcochinase, which is at the low end of the range for related β-glucosidases of pH 5-8 [11]. These residues lie between the catalytic residues and the protein surface, in the cleft where the aglycone is likely to bind, so they might also influence the substrate specificity. A substitution of Ser for Trp may allow for the large rotenoid structure of dalcochinin to fit into this cleft more easily to interact with the many aromatic residues in this region. The smaller structures of linamarin and lotaustralin, the substrates for the white clover enzyme, may require the Trp in this position for tight binding. Ser is also seen in this position in the noncyanogenic β-glucosidase from cassava (Figure 4), so perhaps this enzyme hyrolyzes more bulky substrates as well. Steric considerations may play discriminate in the opposite direction closer to the active site, since dalcochinin is a primary alcohol, linamarin and lotaustralin are tertiary alcohols. More structural and functional studies will also be necessary to elucidate which amino acids may contact the aglycone in the entrance to the catalytic pocket and account for the differences in specificity between the *D. cochinchinensis* enzyme and the white clover enzyme.

The conservation of the cysteine residues with those of the white clover

cyanogenic β-glucosidase [15], suggests they may be structurally important, though they are not conserved in other related enzymes (Figure 4). Only residues 201 and 209 form an internal disulfide bond, so that other cysteines may be involved in disulfide bonds between subunits. Indeed, some, but not all, of the protein migrates at high molecular weight when no β-mercaptoethanol is added during its preparation for SDS-PAGE (data not shown).

The sequence similarity with β-glucosidases that generate noxious compounds suggests that the dalcochinase enzyme may also play some role in defense against herbivores, as does its substrate, which is similar to the respiratory inhibitor rotenone. This role is being investigated by studies of the bioactivity of the substrate. The apparent specificity of expression of enzyme activity in seeds and young plants will also need to be explored further to determine the biological significance of the enzyme. β-Glucosidases have also been implicated in activation of cytokinin growth factors and phosphate stress response [6,7]. Bacterial β-glucosidases have been found to play a role in plant-microbe interactions by activation of plant growth during infection [32,33]. Thai Rosewood is a legume, which might produce glycosylated signal molecules to interact with rhizobia bacteria during early stages of growth. So it is not yet possible to rule out that the enzyme and its substrate have some role in growth or environmental interactions outside of defense against herbivores. Further investigations will help clarify the biological role of the enzyme.

Acknowledgments:

The authors are greatly indebted to Professor Brigitte Wittman-Liebold and Bernd Thiede for help with sequencing and for running the mass-spectroscopy experiment at

the Max Delbruck Centrum f r Molekular Medizin. Dr. Rudee Surarit provided excellent advice and discussion throughout the project. Phannee Sawangareetrakul, Pantipa Subhasitanont, Daranee Chokchaichamnankit, and Kanokporn Boonpuan provided technical assistance in various aspects of protein purification. The work was supported by a grant from Chulabhorn Research Institute. JRKC is supported by grant RSA011/2539 from the Thailand Research Fund.

References:

1. Campo, N., Bako, L., Brzobohaty, B. Feldwisch, J., Zettle, R., Boland, W. and Palme, K. (1993) In: Esen. A. Editor, "β-glucosidases: Biochemistry and Molecular Biology." ACS Symposium Series **533**, American Chemical Society. Washington D.C., USA, 205-213

2. Henrissat, B. (1991) *Biochem. J.* **280**, 309-316.

3. Kuroki, G.W. and Poulton, J.E. (1987) *Arch. Biochem. Biophys.* **255**, 19-26

4. Poulton, J.E. (1993) In: Esen. A. Editor, "β-glucosidases: Biochemistry and Molecular Biology." ACS Symposium Series **533**, American Chemical Society. Washington D.C., USA.

5. McDannell, R., McLean, A.E.M., Hanley, A.B., Heanley, R.K. and Fenwick, G.R. (1988) *Food Chem. Toxicol.* **26**, 59-70

6. Brzobohaty, B., Moore, L., Kristoffersen, P., Bako, L., Campos, N., Schell, J. and Palmet, K. (1993) *Science* **262**, 1051-1053

7. Malboobi MA and Lefebvre DD. (1997) *Plant Mol. Biol.* **34**,157-68

8. Nilsson, K. (1988) *Trends in Biotech.* **6**, 256-264.

9. Welply, J.K. (1989) *Trends Biotech.* **7**, 5-8.

10. Surarit, R. Svasti, J. Srisomsap, C., Suginta, W., Khunyoshyeng, S., Nilwarangkoon, S., Harnsakol, P., and Benjavongkulchai, E. (1995) *J. Sci. Soc. Thailand* **21**, 293-303

11. Srisomsap, C., Svasti, J., Surarit, R., Champattanachai, V. Sawangareetrakul, P., Boonpuan, K., Subhasitanont, P. and Chokechaichamnankit, D. (1996) *J. Biochem.* **119**, 585-590

12. Surarit, R. Chiaba, S. Matsui, H., Svasti, J. and Srisomsap, C. (1996) *Biosci. Biotech. Biochem.* **60**(8), 1265-1268

13. Svasti, J., Srisomsap, C., Techasakul, S. and Surarit, R. (1999) *Phytochemistry* **50**, 739-743

14. Henrissat, B. and Bairoch, A. (1993) *Biochem J.* **293**, 781-783

15. Barrett, T., Suresh, C.G., Tolley, S.P., Dodson, E.J., and Hughes, M.A. (1995) *Structure* **3**, 951-960.

16. Bednar, R.A. and Hadcock, A.L. (1988) *J. Biol. Chem.* **263**, 9582-9588

17. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.* **215**,403-410.

18. Madden, T.L., Tatusov, R.L. and Zhang, J. (1996) *Meth. Enzymol.* **266**, 131-141.

19. Altschul, S.F., Madden, T.L., Schbffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res.* **25**, 3389-3402.

20. Feng, D.F. and Doolittle, R.F. (1996) *Meth. Enzymol.* **266**, 368-382

21. Chomczynski, P. and Sacchi, N. (1987) *Anal. Biochem.* **162**, 156-159

22. Sambrook, J., Fritsch, E.F. and Maniatis, T., (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

23. Guex, N. and Peitsch, M.C. (1997) *Electrophoresis* **18**, 2714-2723.

http://www.expasy.ch/spdbv/mainpage.htm

24. Hays, W.S., Jenison, S.A., Yamada, T., Pustuszyn, A. and Glew, R.H. (1996) *Biochem. J.* **319**, 329-837

25. Burmeister, W.P., Cottaz, S., Driguez, H., Iori, R., Palmieri, S. and Henrissat, B. (1997) *Structure* **5**, 663-675

26. Hofmann K., Bucher P., Falquet L. and Bairoch A. (1999) *Nucleic Acids Res.* **27**, 215-219

27. Bucher P. and Bairoch A. (1994)(In) ISMB-94; *Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology*. Altman R., Brutlag D., Karp P., Lathrop R. and Searls D., Eds., pp. 53-61, AAAIPress, Menlo Park

28. Withers, S.G. and Street, I.P. (1988) *J. Am Chem Soc.* **110**, 8551-8553

29. Keresztessy, Z., Kiss, L. and Hughes, M.A. (1994) *Arch. Biochem. Biophys.* **315**, 323-330

30. Hrmova, M., MacGregor, E.A., Biely, P., Stewart, R.J. and Fincher, G.B. (1998) *J. Biol Chem.* **274** (18), 11134-11143

31. Jenkins, J. Leggio, L.L., Harris, G. and Pickersgill, R. (1995) *FEBS Letters* **362**, 281-285

32. Estruch, J.J. Shell, J. and Spena, A. (1991) *EMBO J.* **10**, 3125-3128.

33. Estruch, J.J., Chriqui, D., Grossmann, K., Schell, J. and Spena, A., *EMBO J.* **10**, 2889-2895

Table I: Oligonucleotides Used in Cloning and Sequencing

A. Degenerate Oligonucleotides Used for RT-PCR:

For.1 5'-GGGATCC AT(T/C/A) GA(T/C) TT(T/C) GC(N) AA(A/G) GA(A/G) GT-3'

Rev.1 5'-GGGAAGC TT (A/G)TA (A/G)TA (A/G)TG (A/G)AA (C/T)TG(G/A)TC-3'

Rev.2 5'-CC(A/G) TT(C/T) TC(N) GT(A/G/T) AT(A/G) TA-3'

B. Oligonucleotides Used in 3' RACE

AP 5'-GGC CAC GCG TCG ACT AGT ACT TTT TTT TTT TTT TTT T-3'

AUAP 5'-GGC CAC GCG TCG ACT AGT AC-3'

For.2 5'-AGG TTC CTC CAT TCA ACC GAA G-3'

For.3 5'-CCA CCA ATA TCC AGA AAA GAT AGC G-3'

C. Oligonucleotides Used in 5' RACE

Oligo-dC-SalI 5'-TTG TCG ACC CCC CCC CCC CCC-3'

Rev3. 5'-AAC ATC TCC GTT GCT TCT ATC CGC-3'

Rev4. 5'-GCT GTC CCA AAA ATG AAA TCT GA-3'

D. Oligonucleotides Used for Sequencing 3' RACE Product and Full cDNA

For4. 5'-AGC CTT GGA GGA TGA GTA CGG T-3'

For5. 5'-TCA AAA TGC TAC CCA GCG ATA TCT-3'

For6. 5'-CGC CAT CTC TTT TAT ATT CGA TAT GC-3'

For7. 5'-GGC AAG ATG GAG CTT ATC AAC G-3'

Rev5. 5'-TTT TGT CTT TTC CTT GAT CAT CG-3'

Rev6. 5'-CGC CAG ACC TAA TTG CAT ATC G-3'

Rev7. 5'-CAT CCA CAT GTG AAG TCA AGA TAT CG-3'

Rev8. 5'-AAT TGG AAF CAA AGA TCC GCA TA-3'

E. Oligonucleotides Used to Amplify Full Coding Region of cDNA.

5'-Term 5'-CTT CCT TTC ATC TCA TGC TTG CA-3'

3'-Term. 5'-AAA GAG AAT ACA ATT CTT TTT GGG CG-3'

Figure 1. Proteolytic Peptide Maps of *Dalbergia cochinchinensis* Pierre β-glucosidase and Sequences of Peptides. Purified and deglycosylated protein was digested with Endopeptidase Lysine C (A) and Trypsin (B) and peptides were purified by reverse-phase HPLC as described in the methods. Peptides from selected peaks were sequenced by automated Edman degradation sequencing and the determined peptide sequences are listed below the appropriate peptide map chromatograph. The peptide sequences are designated by their digestion enzyme (Lys for Endo Lys C or TP for trypsin) and the peak number in the peptide map. Secondary inspection of the sequences of trypsin digest peak 7, indicated two signals with the sequences TP7A and TP7B as indicated. Asparagine was not seen at position 6 of TP7B, but was derived from the nucleic acid sequence.

Figure 2: Map of the Sequence Clones and Peptide Sequences of the β-Glucosidase. The sequence of the entire mRNA (cDNA) is represented by the top band marked mRNA. The primers (small arrows) used for initial reverse transcription and PCR are indicated below this band, followed by the band representing the original PCR product. Below this are the primers designed from this sequence and the other primers used for 3′ and 5′, followed by 3′ RACE and 5′ RACE products. The 5′ term and 3′ term primers used for amplification of the full coding region for sequencing and their amplification product are shown below this. Other primers used for sequencing are shown below this product, along with segments sequenced by direct sequencing (indicated by long, thin arrows). These primers were also used for sequencing the cloned RACE products.

Figure 3: Nucleic Acid and Protein Sequence of the *D. cochinchinensis* β-glucosidase, Genbank accession, AF163097. Peptides are marked under the sequences. Possible glycosylation sites are marked by stars and the site confirmed by mass-spectroscopy and peptide sequencing is in bold. Primer sequences are indicated by bold print in the nucleotide sequence.

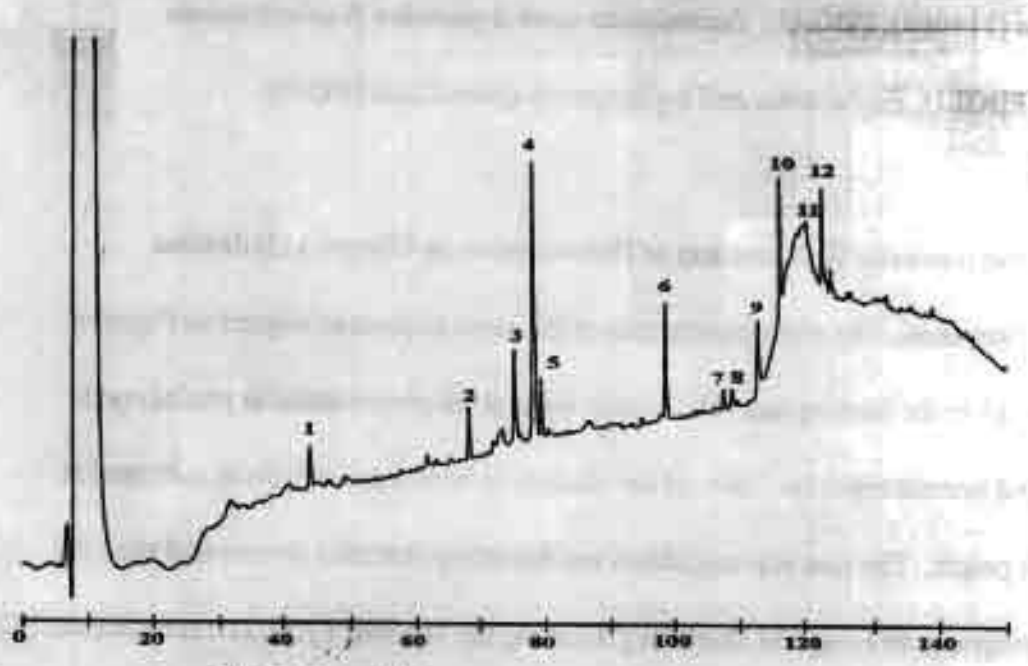
Figure 4: Multiple Sequence Alignment of *D. cochinchinensis* β-glucosidase with related enzymes. The sequence of the mature protein for the *D. cochinchinensis* β-glucosidase is aligned with the mature protein sequences for other β-glucosidase and glycosyl hydrolase family 1 enzymes. Stars indicate residues that are completely conserved among the selected enzyme sequences. The catalytic nucleophile (Glu 395) and catalytic acid/base (Glu 182) are indicated in bold font. The sequences were retrieved from the NCBI nr database using the Entrez server and aligned with the Feng and Doolittle Propack programs[20], using the Dayhoff PAM 250 substitution matrix. The sequences (and nr accessions) are: DCDBGLU, *D. cochinchinensis* β-glucosidase (dalcochinase); 1CBG, white clover cyanogenic β-glucosidase (1311386); BCAHI, black cherry amygdalin hydrolase AH-1(833835); BCPH, black cherry prunasin hydrolase (1236961); NCBGLU, white clover non-cyanogenic β-glucosidase, (114974); CVNCBGLU, cassava non-cyanogenic β-glucosidase (1155090); CVNCBGLU, cassava cyanogenic β-glucosidase (linamarase) (249262); wcncbg1, BBGQ60, barley beta-glucosidase BGQ60 (136216); MYRMB3, white mustard myrosinase 3D (127734); LPHdom3: Human lactase phlorizinhydrolase domain 3 (187053); AGSPBGLU, *Agrobacterium sp.* (strain ATCC 21400) β-

glucosidase (114966); PBGAL, *Lactococcus casei* 6-phospho-β-galactosidase (P14696); PBGLU, *Escherichia coli* 6-phospho-β-glucosidase (24240).

Figure 5. Phylogenetic Relationship of Dalcochinase to Glycosyl Hydrolase Family 1 Enzymes. The phylogenetic tree of the same sequences aligned in Figure 4 is represented in the phenogram. The branch lengths are proportional to phylogenetic distance and percent reproducibility of the clusters in bootstrap analysis is indicated at the branch points. The tree was calculated and bootstrap statistics determined with the Propack programs of Feng and Doolittle [20] using the Dayhoff PAM 250 substitution matrix. The numbers at the branch points are the bootstrap statistics in percent of bootstraps producing the cluster at that branch.

Figure 6. Homology Model of the Dalcochinase 3-Dimensional Structure. The *D. cochinchinensis* dalcochinase sequence (a & c) was modeled on the white clover cyanogenic β-glucosidase 1CBG (b & d) as described in the methods. The overall topology of the dalcochinase and 1CBG structures are shown in a and b, respectively, with the catalytic residues in stick representation with the catalytic nucleophile in cyan and the catalytic acid in green. The cysteine residues are shown in yellow in space-filling representation. The catalytic sites are detailed in c and d which show all residues within 6 angstroms of the catalytic nucleophile and catalytic acid. Acidic residues are shown in red, basic residues in blue, polar in yellow, and nonpolar in gray. The catalytic residues and the amino acids with nonconserved changes between dalcochinase and 1CBG are labeled.

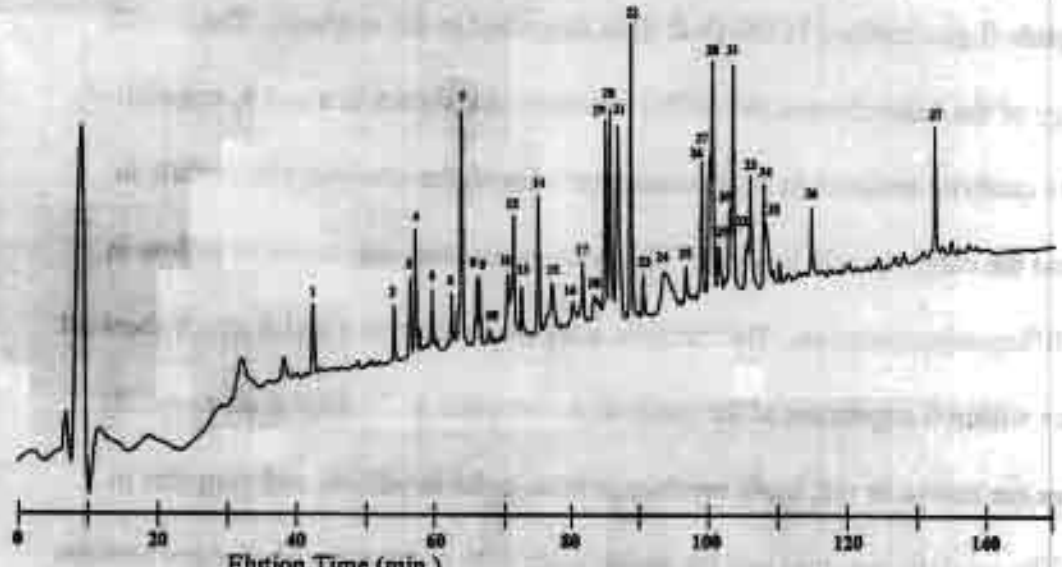
Figure 1 Proteolytic Peptide Maps of *Dalbergia cochinchinensis* Pierre β-glucosidase and Sequences of Peptides.



Elution Time (min.)

A. Endo Lysine C Peptide Map and Determined Peptide Sequences:

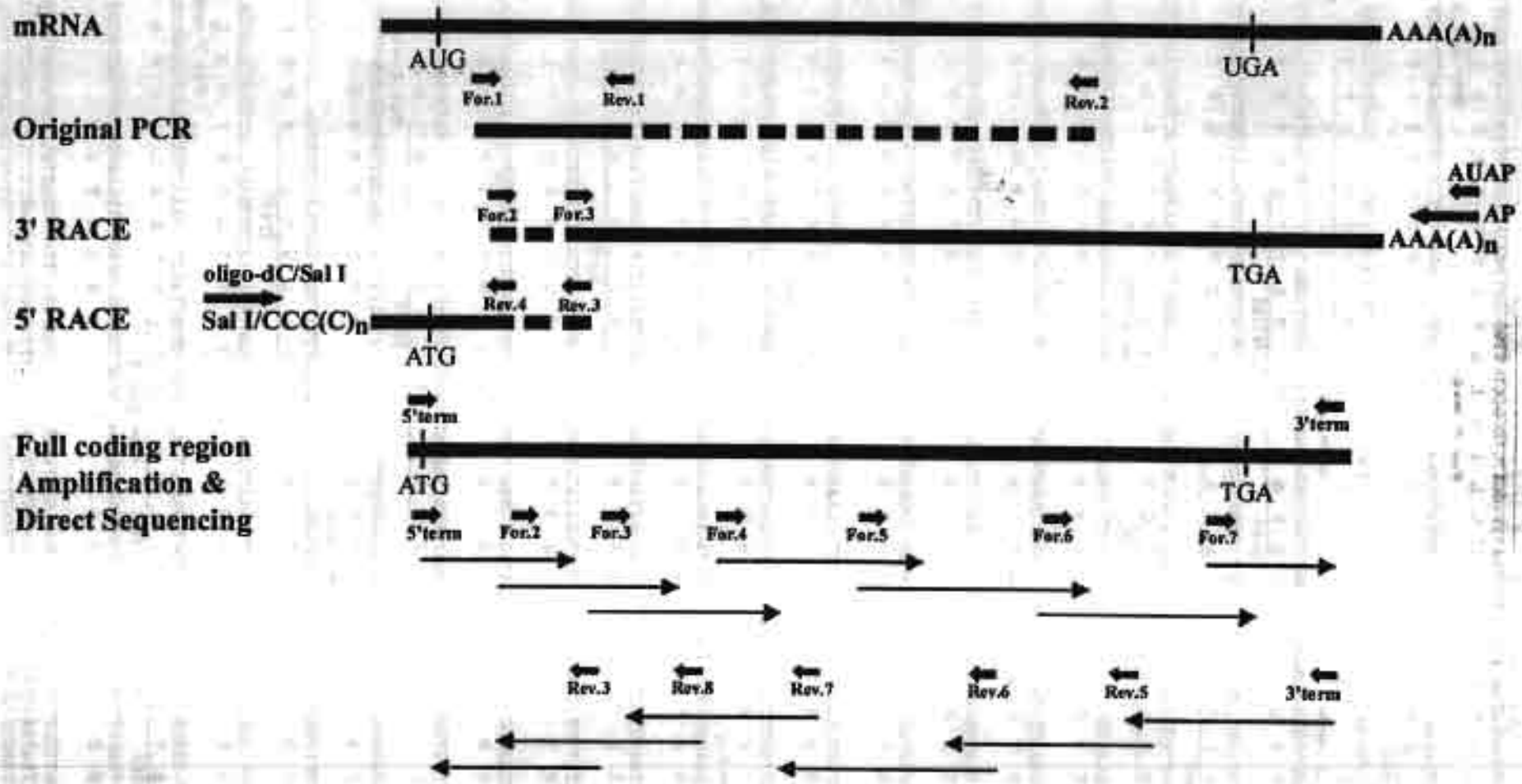
Lys2:	I-D-F-A-K
Lys3:	Y-F-L-A-R-D-Q-X-S-A
Lys4:	I-A-D-X-S-N-G-D-V-A-V-D-Q-H
Lys5:	D-I-A-I-M-L-V-G-X-X-A-V-D-Q
Lys6:	L-S-A-T-X-F-K
Lys9:	Y-N-N-P-L-V-Y-I-T-E-N-G-I
Lys10:	T-S-F-X-F-I-G-L-N-Y-X-T-T-N



Elution Time (min.)

B. Tryptic Peptide Map and Determined Peptide Sequences.

TP7	S-N-G-D-V-A-V-D-Q-F-H-Y-Y-N
Trp7A:	S-N-G-D-V-A-V-D-Q-F-H-R
Trp7B:	V-S-G-G-I-(N)-Q-T-G-V-D-Y-Y-N-(R)
TP9	D-M-N-L-D-A-Y
TP22	L-S-A-T-X-F-K
TP23	V-P-S-I-X-X-N-F-T
TP37	X-F-F-A-X-X-L-L-D


mRNA
AUG
UGA
AAA(A)n
For.1
Rev.1
Rev.2
Original PCR
AUAP
AP
3' RACE
For.2
For.3
AAA(A)n
TGA
oligo-dC/Sal I
5' RACE
Sal I/CCC(C)n
Rev.4
Rev.3
ATG
Full coding region
Amplification &
Direct Sequencing
5'term
3'term
ATG
TGA
5'term
For.2
For.3
For.4
For.5
For.6
For.7
Rev.3
Rev.8
Rev.7
Rev.6
Rev.5
3'term

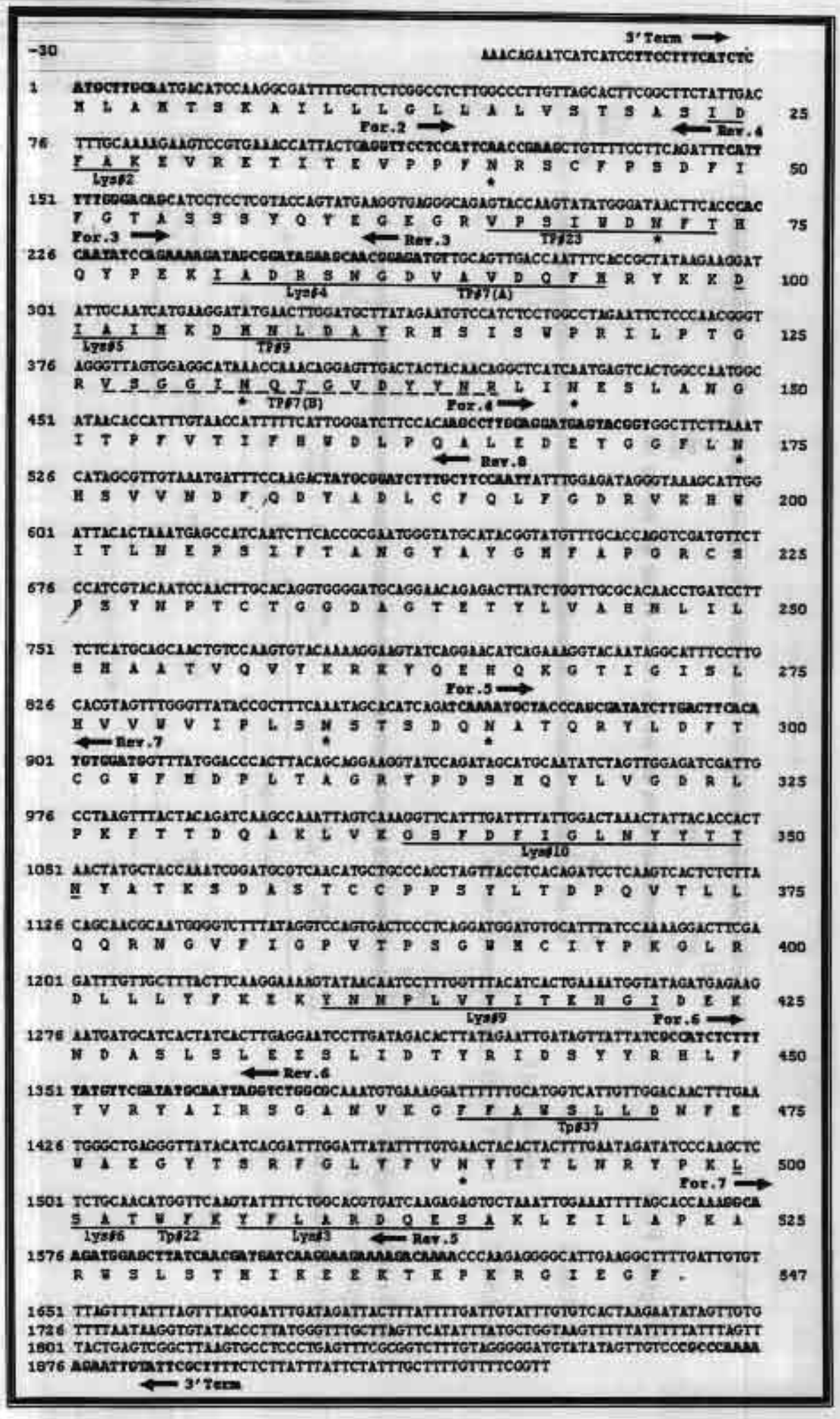
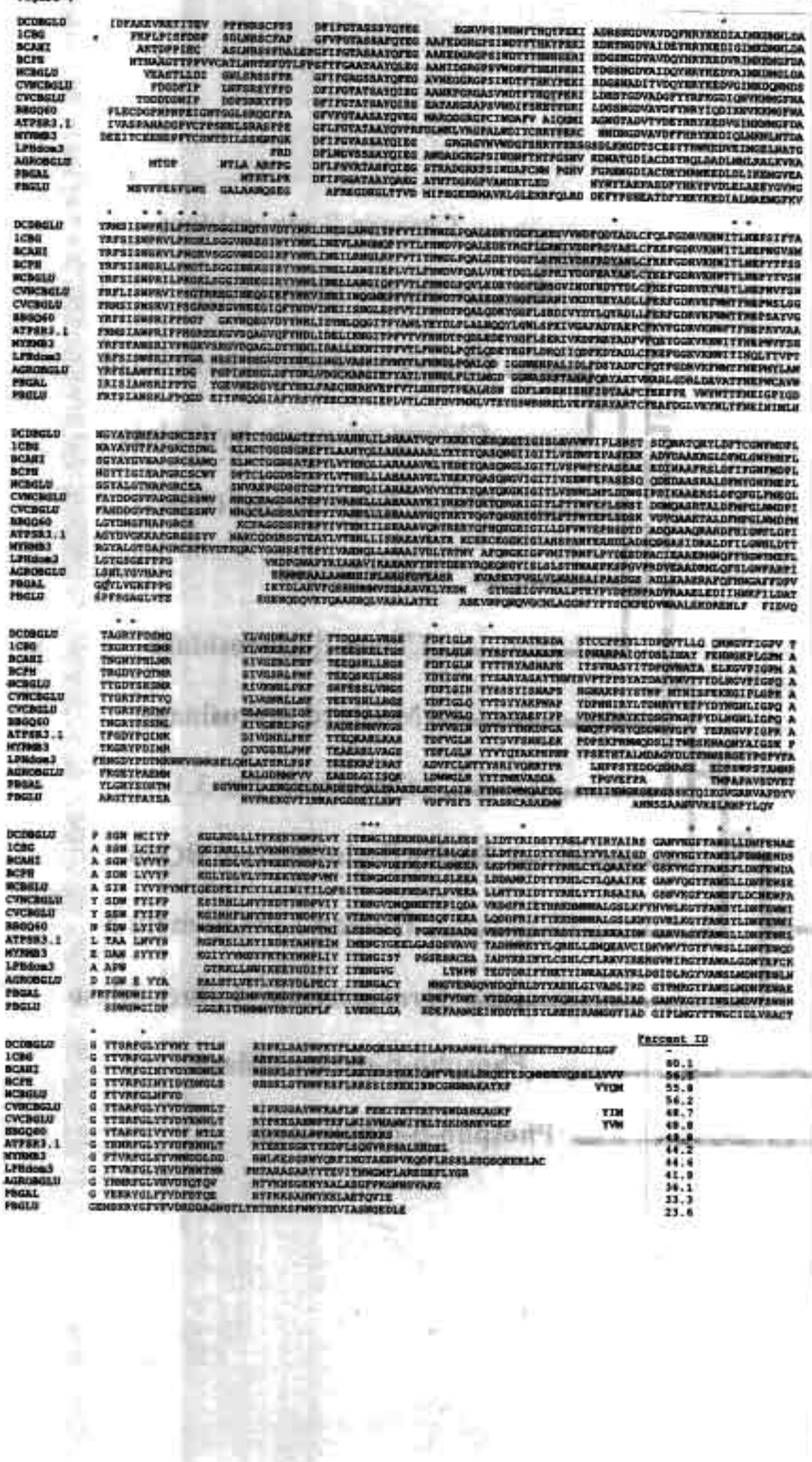
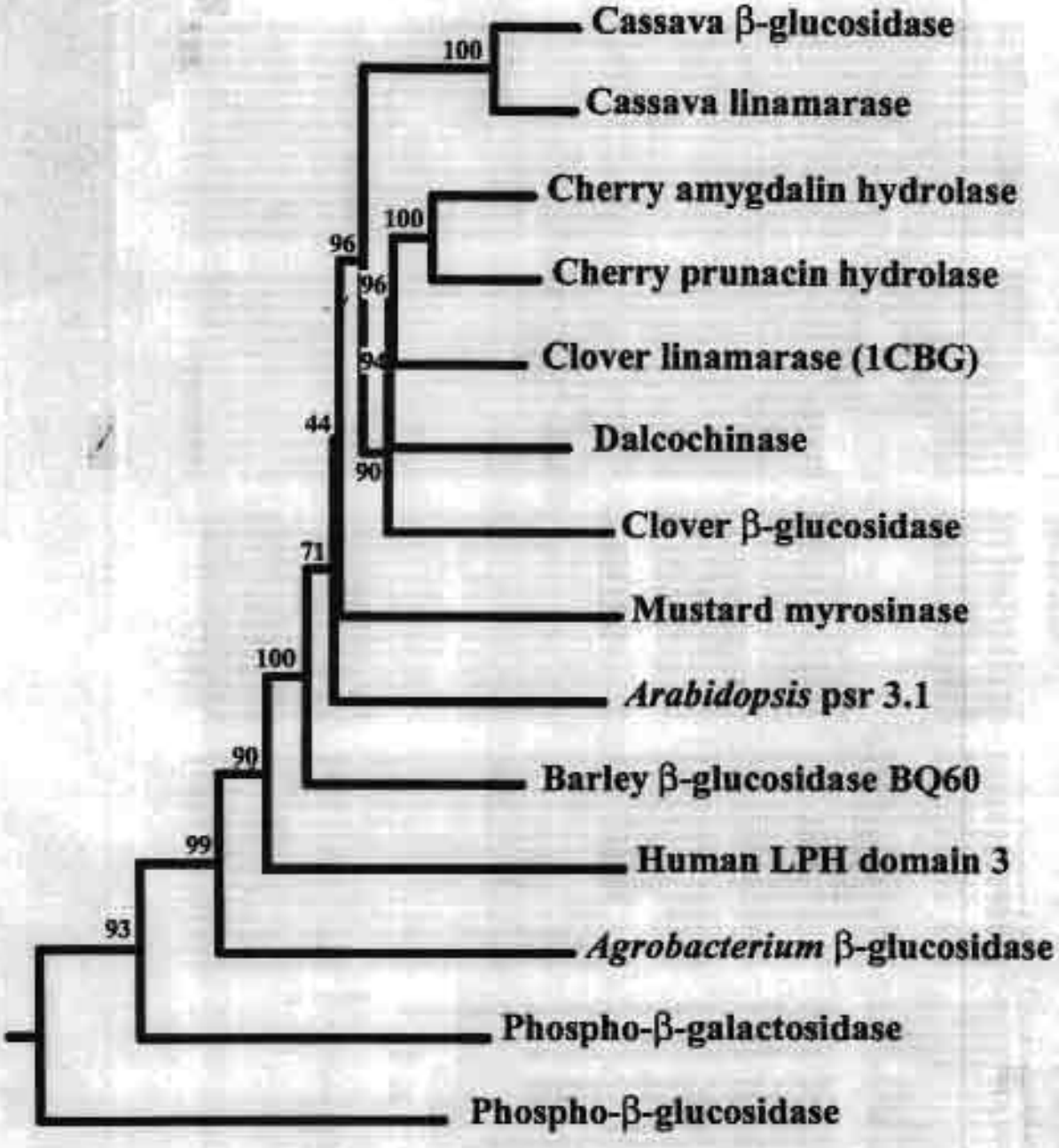
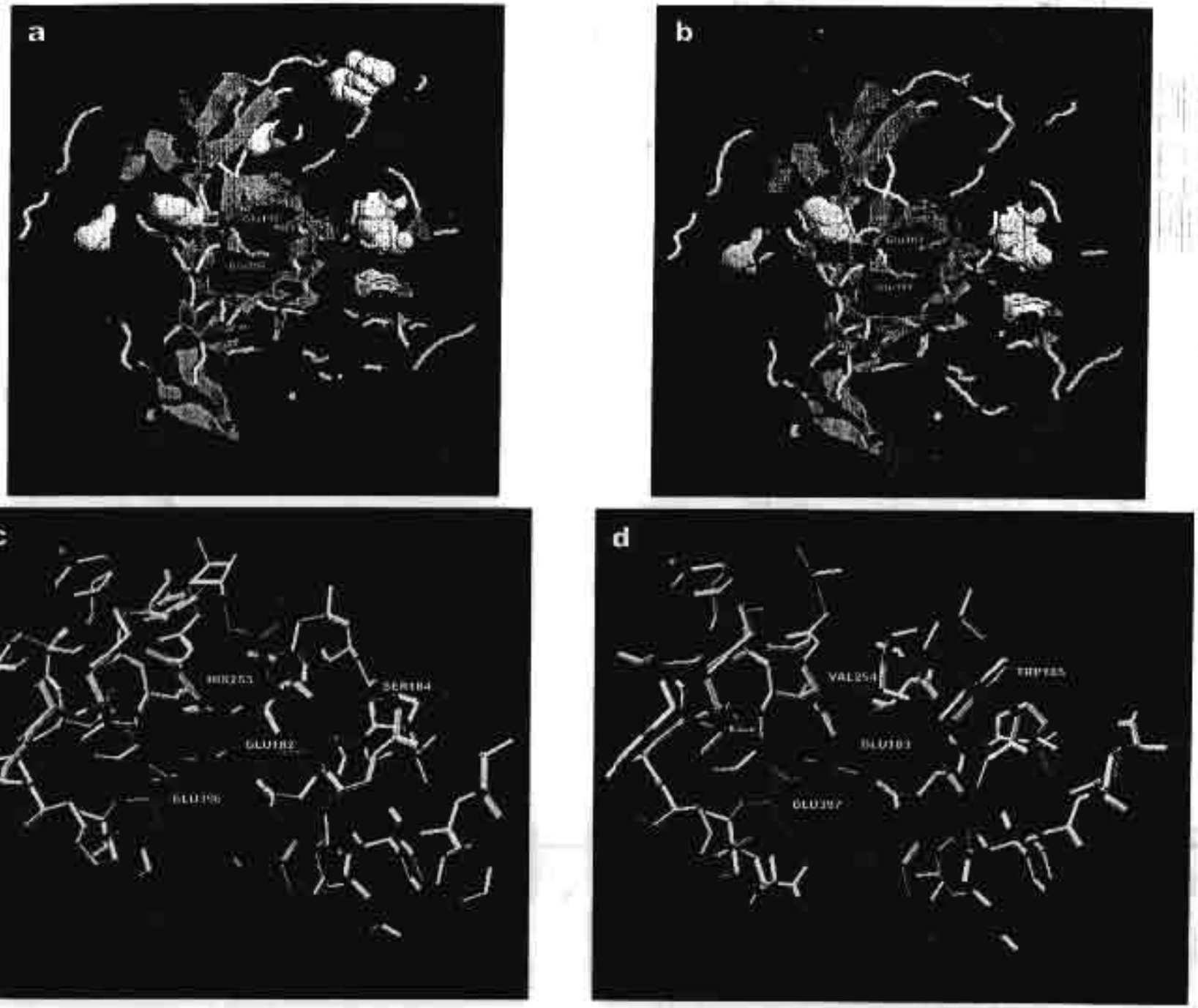

3'Term
-30
AAACAGAATCATCATCCTTCCTTTCATCTC
1 ATGCTTGCAATGACATCCAAGGCGATTTTGCTTCTCGGCCTCTTGGCCCTTGTTAGCACTTCGGCTTCTATTGAC
M L A M T S K A I L L L G L L A L V S T S A S I D 25
For.2
Rev.4
76 TTTGCAAAAGAAGTCCGTGAAACCATTACTCAGGTTCCTCCATTCAACCGAAGCTGTTTTCCTTCAGATTCATT
F A K E V R E T I T E V P P F N R S C F P S D F I 50
Lys#2
151 TTTGGGACAGCATCCTCCTCGTACCAGTATGAAGGTGAGGGCAGAGTACCAAGTATATGGGATAACTTCACCCAC
F G T A S S S Y Q Y E G E G R V P S I W D N F T H 75
For.3
Rev.3
TP#23
226 CAATATCCAGAAAAGATAGCGGATAGAAGCAACGGAGATGTTGCAGTTGACCAATTTCACCGCTATAAGAAGGAT
Q Y P E K I A D R S N G D V A V D Q F H R Y K K D 100
Lys#4
TP#7(A)
301 ATTGCAATCATGAAGGATATGAACTTGGATGCTTATAGAATGTCCATCTCCTGGCCTAGAATTCTCCCAACGGGT
I A I M K D M N L D A Y R M S I S W P R I L P T G 125
Lys#5
TP#9
376 AGGGTTAGTGGAGGCATAAACCAAACAGGAGTTGACTACTACAACAGGCTCATCAATGAGTCACTGGCCAATGGC
R V S G G I N Q T G V D Y Y N R L I N E S L A N G 150
TP#7(B)
For.4
451 ATAACACCATTTGTAACCATTTTTCATTGGGATCTTCCACAAGCCTTGCAGGATGAGTACGGTGGCTTCTTAAAT
I T P F V T I F H W D L P Q A L E D E Y G G F L N 175
Rev.8
526 CATAGCGTTGTAAATGATTTCCAAGACTATGCGGATCTTTGCTTCCAATTATTTGGAGATAGGGTAAAGCATTGG
H S V V N D F Q D Y A D L C F Q L F G D R V K H W 200
601 ATTACACTAAATGAGCCATCAATCTTCACCGCGAATGGGTATGCATACGGTATGTTTGCACCAGGTCGATGTTCT
I T L N E P S I F T A N G Y A Y G M F A P G R C S 225
676 CCATCGTACAATCCAACTTGCACAGGTGGGGATGCAGGAACAGAGACTTATCTGGTTGCGCACAACCTGATCCTT
P S Y N P T C T G G D A G T E T Y L V A H N L I L 250
751 TCTCATGCAGCAACTGTCCAAGTGTACAAAAGGAAGTATCAGGAACATCAGAAAGGTACAATAGGCATTTCCTTG
S H A A T V Q V Y K R K Y Q E H Q K G T I G I S L 275
For.5
826 CACGTAGTTTGGGTTATACCGCTTTCAAATAGCACATCAGATCAAAATGCTACCCAGCGATATCTTGACTTCACA
H V V W V I P L S N S T S D Q N A T Q R Y L D F T 300
Rev.7
901 TGTGGATGGTTTATGGACCCACTTACAGCAGGAAGGTATCCAGATAGCATGCAATATCTAGTTGGAGATCGATTG
C G W F M D P L T A G K Y P D S M Q Y L V G D R L 325
976 CCTAAGTTTACTACAGATCAAGCCAAATTAGTCAAAGGTTCATTTGATTTTATTGGACTAAACTATTACACCACT
P K F T T D Q A K L V K G S F D F I G L N Y Y T T 350
Lys#10
1051 AACTATGCTACCAAATCGGATGCGTCAACATGCTGCCCACCTAGTTACCTCACAGATCCTCAAGTCACTCTCTTA
N Y A T K S D A S T C C P P S Y L T D P Q V T L L 375
1126 CAGCAACGCAATGGGGTCTTTATAGGTCCAGTGACTCCCTCAGGATGGATGTGCATTTATCCAAAAGGACTTCGA
Q Q R N G V F I G P V T P S G W M C I Y P K G L R 400
1201 GATTTGTTGCTTTACTTCAAGGAAAAGTATAACAATCCTTTGGTTTACATCACTGAAAATGGTATAGATGAGAAG
D L L L Y F K E K Y N N P L V Y I T K N G I D E K 425
Lys#9
For.6
1276 AATGATGCATCACTATCACTTGAGGAATCCTTGATAGACACTTATAGAATTGATAGTTATTATCGCCATCTCTTT
N D A S L S L E E S L I D T Y R I D S Y Y R H L F 450
Rev.6
1351 TATGTTCGATATGCAATTAGGTCTGGCGCAAATGTGAAAGGATTTTTTGCATGGTCATTGTTGGACAACTTTGAA
Y V R Y A I R S G A N V K G F F A W S L L D N F E 475
Tp#37
1426 TGGGCTGAGGGTTATACATCACGATTTGGATTATATTTTGTGAACTACACTACTTTGAATAGATATCCCAAGCTC
W A E G Y Y T S R F G L Y F V N Y T T L N R Y P K L 500
For.7
1501 TCTGCAACATGGTTCAAGTATTTTCTGGCACGTGATCAAGAGAGTGCTAAATTGGAAATTTTAGCACCAAAGGCA
S A T W F K Y F L A R D Q E S A K L E I L A P K A 525
Lys#6
Tp#22
Lys#3
Rev.5
1576 AGATGGAGCTTATCAACGATGATCAAGGAAGAAAAGCAAAACCCAAGAGGGGCATTGAAGGCTTTTGATTGTGT
R W S L S T N I K E E K T K P K R G I E G F . 547
1651 TTAGTTTATTTAGTTTATGGATTTGATAGATTACTTTATTTTGATTGTATTTGTGTCACTAAGAATATAGTTGTG
1726 TTTTAATAAGGTGTATACCCTTATGGGTTTGCTTAGTTCATATTTATGCTGGTAAGTTTTTATTTTTATTTAGTT
1801 TACTGAGTCGGCTTAAGTGCCTCCCTGAGTTTCGCGGTCTTTGTAGGGGGATGTATATAGTTGTCCCGCCCAAAA
1876 AGAATTGTATTCGCTTTTCTCTTATTTATTCTATTTGCTTTTGTTTTCGGTT
3'Term

Figure 4



	Percent ID
DCDBGLU	-
1CBG	80.1
BCAXI	56.2
BCFH	55.8
HCBGLU	56.2
CVHCBGLU	48.7
CVCBGLU	48.8
BBGQ60	48.8
ATPSR3.1	44.2
MYRMB3	44.6
LPHdom3	41.9
AGROBGLU	36.1
PBGAL	33.3
PBGLU	23.6


Cassava β-glucosidase
Cassava linamarase
Cherry amygdalin hydrolase
Cherry prunacin hydrolase
Clover linamarase (1CBG)
Dalcochinase
Clover β-glucosidase
Mustard myrosinase
Arabidopsis psr 3.1
Barley β-glucosidase BQ60
Human LPH domain 3
Agrobacterium β-glucosidase
Phospho-β-galactosidase
Phospho-β-glucosidase
100
100
96
96
94
44
90
71
100
90
99
93


a
b
c
d
SER184
GLU182
VAL254
TRP185
GLU183