Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174

Sequence, Structural, Functional, and Phylogenetic Analyses of Three Glycosidase Families

Submitted 01/05/98
(communicated by Ernest Beutler, M.D., 03/19/98)

I. S. Mian

ABSTRACT: Glycosidases, which cleave the glycosidic bond between a carbohydrate and another moiety, have been classified into over 63 families. Here, a variety of computational techniques have been employed to examine three families important in normal and abnormal pathology with the aim of developing a framework for future homology modeling, experimental and other studies. Family 1 includes bacterial and archaeal enzymes as well as lactase phlorizin-hydrolase and klotho, glycosidases implicated in disaccharide intolerance II and aging respectively. A statistical model, a hidden Markov model (HMM), for the family 1 glycosidase domain was trained and used as the basis for comparative examination of the conserved and variable sequence and structural features as well as the phylogenetic relationships between family members. Although the structures of four family 1 glycosidases have been determined, this is the first comparative examination of all these enzymes. Aspects that are unique to specific members or subfamilies (substrate binding loops) as well those common to all members (a $(\beta/\alpha)_8$ barrel fold) have been defined. Active site residues in some domains in klotho and lactase-phlorizin hydrolases differ from other members and in one instance may bind but not cleave substrate. The four invariant and most highly conserved residues are not residues implicated in catalysis and/or substrate binding. Of these, a histidine may be involved in transition state stabilization. Glucosylceramidase (family 30) and galactosylceramidase (family 59) are mutated in the lysosomal storage disorders Gaucher disease and Krabbe disease, respectively. HMM-based analysis, structure prediction studies and examination of disease mutations reveal a glycosidase domain common to these two families that also occurs in some bacterial glycosidases. Similarities in the reactions catalyzed by families 30 and 59 are reflected in the presence of a structurally and functionally related $(\beta/\alpha)_8$ barrel fold related to that in family 1.

Keywords: hidden Markov model, Klotho, aging, Gaucher disease, Krabbe disease, protein evolution, TIM barrel, glycosidase

INTRODUCTION

Glycosyl hydrolases (glycosidases) are a diverse group of evolutionarily conserved enzymes which cleave the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non-carbohydrate (aglycone) moiety (reviewed in (1-3)). Enzymatic hydrolysis occurs via a two step mechanism: formation of a glycosyl-enzyme with concomitant aglycone departure followed by hydrolysis of the glycosyl-enzyme by a water molecule. The reaction leads to either retention or inversion of stereochemistry. Based on their primary sequences, glycosidases have been classified into over 63 families grouped into five clans or superfamilies (1). The three families of interest in this work (families 1, 30, and 59) contain members which have roles in both normal and abnormal pathology: aging, disaccharide intolerance II and two lysosomal storage disorders (Gaucher disease and Krabbe disease). Since these enzymes are known or putative ceramidases, they may be important in apoptosis. Here, a variety of computational techniques have been employed to

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA.
Reprint requests to: I.S. Mian, Life Sciences Division (Mail Stop 29-100), Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720. E-mail: SMian@lbl.gov; Tel: 510 486-6216; Fax: 510 486-6949.

1079-9796/98 $25.00

analyze these glycosidase at the molecular level in order to provide insights into their structures and mechanisms of action and thus a framework for further studies.

Family 1 is composed of β-glucosidases, β-galactosidases, 6-phospho-β-galactosidases, 6-phospho-β-glucosidases, myrosinases (thioglucosidase, sinigrinase) and lactase-phlorizin hydrolases (β-galactosidase, glycosylceramidase). The latter is an intestinal, cell-bound extracellular enzyme involved in disaccharide intolerance II (MIM (4) 22300). The newest member, Klotho, is an extracellular protein composed of two internal repeats (KL1, KL2) and a transmembrane region at its carboxy terminus (5). Each repeat exhibits 20-40% similarity to β-glucosidases from bacteria, plants and mammalian lactase-phlorizin hydrolase (5) and is related to archaeal lactases (6). Mutation of *Mus musculus* Klotho leads to a syndrome with features resembling ageing (5). Gaucher-like disease (MIM 230800, 230900 and 231000) is an inherited deficiency of glucosylceramidase (family 30) that is an imperfect match to the Klotho phenotype. The several forms of Gaucher disease are cerebroside lipidoses and have been diagnosed from the first week of life up to 86 years. Krabbe disease (MIM 245200) is a rare autosomal recessive demyelinating neurodegenerative disorder caused by reduced activity of galactosylceramidase (family 59). Although most patients have the severe infantile form, late-onset forms have been described.

Glycosidases belonging to family 1 have been the most extensively studied. Oligomerization is a common characteristic with some being tetramers (7, 8) and others being dimers (9, 10). Although myrosinases are *S*-glycosidases, they share extensive similarity with the *O*-glycosidase members of this family. Structure prediction (11-13) and crystallographic studies (7, 10, 14, 15) of this family indicate that it has the classic $(\beta/\alpha)_8$ barrel fold first observed in the structure of triose phosphate isomerase (16). The glycosidases whose structures have been determined are cyanogenic β-glucosidase (linamarase) from white clover *Trifolium repens* (10), 6-phospho-β-galactosidase from the mesophilic bacterium *Lactococcus lactis* (14), β-glycosidase from the hyperthermophilic archaeon *Sulfolobus solfataricus* (7) and myrosinase from *Sinapis alba* (15). Glycosylation is believed to involve nucleophilic attack by a critical, conserved glutamic acid (Glu) residue. The water molecule that mediates deglycosylation has been believed to be activated by a second conserved Glu acting as a general base. However, recent studies of myrosinase (15) in which this Glu is replaced by glutamine (Gln) suggest that hydrolysis of the glycosyl-enzyme intermediate occurs via the precise positioning of a nucleophilic water molecule and not by a general base activation of water.

The current study refines and extends preliminary work on KL1/KL2 and family 1 glycosidases (6) that employed the statistical modeling method known as hidden Markov models (HMMs). Profile-based HMMs of the type used here (17-20) can characterize the primary sequence features of a family, generate a multiple sequence alignment, identify new members (database searches) and serve as the basis for structure prediction and phylogenetic studies (see for example (21-32)). Examination of the sequence, structural, functional and phylogenetic features of the family 1 glycosidase domain provides information on the family as a whole as well as specific members, notably Klotho and lactase phlorizin hydrolase. These data were used subsequently to make inferences about the proteins mutated in Gaucher and Krabbe diseases, enzymes for which there is a paucity of information at the molecular level. An HMM trained for a glycosidase domain predicted to be common to families 30 and 59 and some bacterial glycosidases, structure prediction studies, and analysis of Gaucher and Krabbe disease mutations yield insights into the structure, function and active site of this family 30/50 glycosidase domain.

METHODS

Statistical Modeling: Hidden Markov Model

A more detailed description of HMMs as used in modeling families of related sequences can be

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174

found elsewhere (17, 33-35) so only a summary is provided. Such HMMs consist of a sequence of nodes corresponding to the columns in a multiple sequence alignment of the family and can be viewed as profiles recast in a probabilistic framework. A match state corresponds to the consensus position in an alignment, an insert state permits insertions relative to the consensus and a delete state allows consensus positions to be skipped. Training an HMM (estimating its parameters) involves creating a stochastic model representing the family used to train it (the training set) by describing transitions into a match, delete or insert state and the occurrence of a given residue in a particular match or insert state. Generating a multiple sequence alignment for a family involves aligning each member to the HMM rather than to other members. A database search consists of scoring each sequence in a database against the model and evaluating the significance of the resultant score.

HMM creation, training and use were performed with v2.0 of the SAM (Sequence Alignment and Modeling Software System) suite (17, 18) running on a MASPAR MP-2204 with a DEC Alpha 3000/300X frontend at the University of California Santa Cruz (UCSC). To improve the ability of the HMM to generalize, Dirichlet mixture priors (19, 20) were employed. Free Insertion Modules (FIMs) were utilized to allow an arbitrary number of insertions at either end of the HMM to accommodate domains that occurred within larger sequences. Sequencing weighting was used to ameliorate the problem of overrepresentation of some sequences.

In previous work (21-27), remote homologues were identified by an iterative scanning strategy: sequences found in one round of HMM searching were added to the training set and the expanded set was used to retrain the HMM for the next round of searching. This procedure was repeated until convergence. This approach to database searching using HMMs is similar to that obtained by use of PSI-BLAST (36), a recent extension of BLAST (37). PSI-BLAST performs an initial gapped BLAST search of the database. In subsequent iterations, statistically significant alignments from the previous search are employed to construct a position-specific score matrix for use in the next round of searching in place of the query and standard amino acid substitution matrix. PSI-BLAST converges and stops if all sequences found at a particular round below a threshold value were already in the model at the beginning of the round.

The SAM program hmm score and an initial HMM trained using 10 sequences (6) were used for HMM database searches by calculating log-odds scores (38, 39) for all sequences in a non-redundant protein database obtained from the NCI (40) and updated weekly at UCSC. The log-odds score for a sequence measures how much more likely it is to have been generated by the HMM as opposed to a competing NULL model consisting of a simple FIM loop (39). The level of significance σ is related to a log-odds score d as follows: $\sigma \leq Nz^{-d}$ (N is the database size and z is the logarithm base) (39). The PSI-BLAST parameter E estimates the statistical significance of a hit by specifying the number of hits with a given score that are expected by chance in a search of a database of given size. The default expected number of false positives E is 0.01 (0.01 matches with a given score would be expected purely by chance). The SAM NULL model, assumed to be a reasonably accurate description of the space the sequences are drawn from, is unlikely to be a good model for the score distribution of all "random" sequences. Hence, although SAM σ values only approximate PSI-BLAST E values ($\sigma \approx E$), the two are comparable and $\sigma = 0.01$ certainly indicates significance in spite of this significance level being pessimistic.

Taking into account the number of sequences in the database searched (~272,000 different proteins in November 1997), a significant log-odds score is considered to be 22.7, the value at which $\sigma = 0.01$. Log-odds scores higher than this value denote fewer expected false positives. The approach employed here emphasizes training an HMM that discriminates between training and non-training set sequences, *i.e.*, one in which the gap in log-odds scores between the lowest scoring

training set sequence and the highest scoring non-training set (database) sequence is relatively large (usually greater than 5.0) and the absolute log-odds score for the lowest training set sequence is greater than 22.7. In addition, efforts were made to ensure that training resulted in an HMM capable of yielding an alignment such that known enzymatic elements aligned.

A database search with the initial HMM revealed a number of sequences with log-odds scores higher than or close to that of the lowest scoring training set sequence. The alignment of sequences with log-odds scores greater than 22.7 was examined and those which possessed regions conserved in the initial HMM were retained and added to the training set. The HMM was then retrained with this expanded training set. Further rounds of "search, align and retrain" revealed fewer and fewer new sequences with the domain. The gap in log-odds scores between training set and non-training set sequences remained relatively constant. At this point (November 1997) and after approximately 10 iterations, a final HMM was trained and used for subsequent studies.

Phylogenetic Analysis

An HMM-generated alignment of the training set containing only match and delete states was utilized for phylogenetic studies. Insert states are not modeled by an HMM because the regions in a sequence they represent are the most divergent parts of the molecules and are likely to be sources of systematic error in phylogenetic analysis. The MOLPHY suite uses a probabilistic procedure for inferring phylogenetic relationships (41, 42). A number of initial trees were generated using the default JTT model for amino acid substitutions and v2.3 of the program. A maximum likelihood distance matrix was calculated and employed to infer a number of approximate trees with NJdist, a neighbor-joining method. The Star Decomposition algorithm was used to calculate another tree (not the maximum likelihood tree). Starting from these initial trees, repeated local rearrangements were employed to search for better tree topologies. Among these final trees, the one with the highest likelihood was selected. Approximate bootstrap probabilities were computed using the RELL method.

RESULTS

Figures showing multiple sequence alignments, phylogenetic trees and ribbon diagrams of molecules were produced using ALSCRIPT (43), Treetool (44) and MOLMOL (45), respectively.

Family 1 Glycosidase Domain: Klotho, Lactase Phlorizin Hydrolase

Hidden Markov model. Table 1 lists the sequences that comprised the final training set (very close homologues are not shown). Of ~273,000 sequences searched using the final HMM, the lowest scoring training set sequence had a log-odds score of 138.5. Only training set sequences and close homologues had scores above this value. The next highest scoring sequences were fragments of *Sinapis alba* myrosinases (44.5-63.1); all other database sequences had scores below 15.4. Thus, sequences with log-odds scores greater than 138.0 are classified as possessing a family 1 glycosidase domain. PSI-BLAST searches with a few randomly selected divergent domains (data not shown) yielded the same set of proteins (those known to belong to this family). The large gap in log-odds scores between training set and non-training set sequences suggests that further generalization of the HMM is required to detect more distantly related family members.

Phylogenetic analysis. Figure 1 shows the phylogenetic tree for the domain. There are 5 major subfamilies: A and E contain bacterial sequences; B plant sequences; C non-plant eucaryotic sequences and D archaeal and bacterial sequences. The two domains in Klotho (Hs_KL1, Hs_KL2; subfamily C) are most similar to the four

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174

Table 1. Family 1 Glycosidases Used in This Study Grouped According to Their Origin

Bacteria		
Agrobacterium ATCC 21400	As_BGLS	β-glucosidase [BGLS_AGRSP]
Bacillus subtilis	Bs_BGL2	probable β-glucosidase [BGL2_BACSU]
	Bs_BGLA	6-phospho-β-glucosidase [BGLA_BACSU]
	Bs_BGLH	β-glucosidase [BGL1_BACSU]
	Bs_YDHP	β-glucosidase [D88802]
Bacillus circulans	Bc_BGLA	β-glucosidase [BGLA_BACCI]
Bacillus polymyxa	Bp_BGLA	β-glucosidase A [BGLA_BACPO]
	Bp_BGLB	β-glucosidase B [BGLB_BACPO]
Bifidobacterium breve	Bb_CLB	β-glucosidase [JC5137]
Caldocellum saccharolyticum	Cs_BGLS	β-glucosidase A [BGLS_CALSA]
Clostridium longisporum	Cl_ABGA	6-phospho-β-glucosidase [ABGA_CLOLO]
Clostridium thermocellum	Ct_BGLA	β-glucosidase A [BGLA_CLOTM]
Erwinia chrysanthemi	Eh_ARBB	6-phospho-β-glucosidase [ARBB_ERWCH]
Erwinia herbicola	Eh_BGLA	β-glucosidase A [BGLA_ERWHE]
Escherichia coli	Ec_ASCB	6-phospho-β-glucosidase [ASCB_ECOLI]
	Ec_BGLA	6-phospho-β-glucosidase [BGLA_ECOLI]
	Ec_BGLB	6-phospho-β-glucosidase [BGLB_ECOLI]
Fusobacterium mortiferum	Fm_PBGA	6-phospho-β-glucosidase [FMU81184]
Klebsiella oxytoca	Ko_CASB	phospho-cellobiase [CASB_KLEOX]
Lactobacillus gasseri	Lg_PBG1	phospho-β-galactosidase 1 [AB003927]
Lactobacillus casei	Lc_LACG	6-phospho-β-glucosidase [LACG_LACCA]
Lactococcus lactis	Ll_4PBG‡	6-phospho-β-galactosidase [4PBG]
Microbispora bispora	Mb_BGLB	thermostable β-glucosidase B [BGLB_MICBI]
Staphylococcus aureus	Su_LACG	6-phospho-β-galactosidase [LACG_STAAU]
Streptococcus mutans	Sm_LACG	6-phospho-β-glucosidase [LACG_STRMU]
Streptomyces QM-B814	Sq_BGL	β-glucosidase [S45675]
Streptomyces rochei	St_BGL	β-glucosidase [S35958]
Thermoanaerobacter brockii	Tb_CGLT	β-xylo-glucosidase [TBZ56279]
Thermotoga maritima	Tm_BGLA	β-glucosidase A [BGLA_THEMA]

Archaea		
Pyrococcus furiosus	Pf_BMNA	β-mannosidase [PFU60214]
	Pf_CELB	β-glucosidase celB [PFU37557]
Sulfolobus shibatae	Sh_BGAL	β-glucosidase (lactase) [BGAL_SULSH]
Sulfolobus solfataricus	Ss_1GOW‡	β-galactosidase [1GOW]
	Ss_BGAS	β-galactosidase (lactase) [BGAL_SULSO]
Thermococcus AL662	Ts_BGLT	β-glucosidase [TSAL6BGLT]

domains in lactase-pholorizin hydrolase. The domains in lactase-pholorizin hydrolase are proposed to have evolved by two cycles of partial gene duplication (46). Thus, the domains in Klotho may have evolved via a duplication event involving the same or a closely related ancestral gene. Also in subfamily C are a yeast exported, cell-associated extracellular β-glucosidase active against cellobiose and all soluble cellodextrins (Cw_BG2) (47, 48), a guinea pig liver cytosolic β-glucosidase with a broad specificity for sugars and a preference for hydrophobic aglycones (Cp_BGL) (49), and a worm glycosidase (Ce_BGL). The four domains of known structure fall into subfamilies A (Ll_4PBG), B (Tr_1CBG, Sa_2MYR) and D (Ss_1GOW). Since

Table 1. (Continued)

Eucarya		
Arabidopsis thaliana	At_MYRO	myrosinase [MYRO_ARATH]
	At_PSR31	β-glucosidase [ATU72153]
Avena sativa	As_BGA	avenacosidase (β-glucosidase) [S50756]
	As_BGA2	β-glucosidase [S43128]
Brassica napus	Bn_BGL1	β-glucosidase [S52771]
	Bn_BGL2	myrosinase [S39549]
	Bn_MYR	myrosinase [S39550]
	Bn_MYRO	myrosinase [MYRO_BRANA]
Candida wickerhamii	Cw_BG2	β-glucosidase [CWU13672]
Cavia porcellus	Cp_BGL	β-glucosidase [CPU50545]
Caenorhabditis elegans	Ce_BGL	similar to lactase-phlorizin hydrolase and β-glucosidases [CELC50F7]
Costus speciosus	Cs_F26G	furostanol glycoside 26-O-β-glucosidase [CSAF26G]
Homo sapiens	Hs_KL	Klotho [AB005142]
	Hs_LPH	lactase-phlorizin hydrolase [LPH_HUMAN]
Hordeum vulgare	Hv_BGQ60	β-glucosidase BGQ60 [A57512]
Manihot esculenta	Me_BLGA	β-glucosidase [MEBGLA]
	Me_LIN	linamarase [MEU95298]
	Me_LIN2	linamarase [S23940]
Oryctolagus cuniculus	Oc_LPH	lactase-phlorizin hydrolase [LPH_RABIT]
Prunus avium	Pa_BGL	β-glucosidase [PAU39228]
Prunus serotina	Ps_AHI	amygdalin hydrolase isoform AH I [PSU26025]
	Ps_PH	cyanogenic prunasin hydrolase [PSU50201]
Rattus norvegicus	Rn_LPH	lactase-phlorizin hydrolase [JS0610]
Sinapis alba	Sa_2MYR‡	myrosinase [2MYR]
	Sa_MYR3	myrosinase MB3 [MYR3_SINAL]
Trifolium repens	Tr_1CBG‡	cyanogenic β-glucosidase [1CBG]
	Tr_BGLS	non-cyanogenic β-glucosidase [BGLS_TRIRP]
Zea mays	Zm_GLU1	β-glucosidase [BGLC_MAIZE]
	Zm_GLU2	β-D-glucosidase [ZMU44087]

The name of the organism, sequence abbreviation and the enzyme are listed. ‡ denotes enzymes whose three-dimensional structures have been solved and whose amino acid sequence are taken from the PDB entries. Databank codes are given in square parenthesis.

Sa_2MYR and Ss_1GOW are likely to be the most and least similar, respectively, to the domains in Klotho, subsequent structural analyses will focus largely on these two enzymes.

Sequence and structural features. An HMM-generated multiple sequence alignment of the training set and the lengths and locations of β-strands and α-helices in domains of known structure were used to infer the secondary structure elements likely to be present in all domains. These results together with the positions most likely to be important for structure and/or function are shown in Figure 2. Although family members range in length from ~380 (Rn_LPH1, Oc_LPH1, Hs_LPH1) to 609 (Cw_BG2) residues, the 356 nodes in the current HMM provide an estimate for the number of positions likely to be common to all sequences that possess the domain. Differences in length can be attributed largely to variations in the size of two loop regions labeled L1 and L2. Although the first domain in lactase-phlorizin hydrolases is the most divergent in terms of primary sequence (Figure 1), it appears to represent the minimal domain. Only 10% of the positions are conserved across all the domains (37/

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174

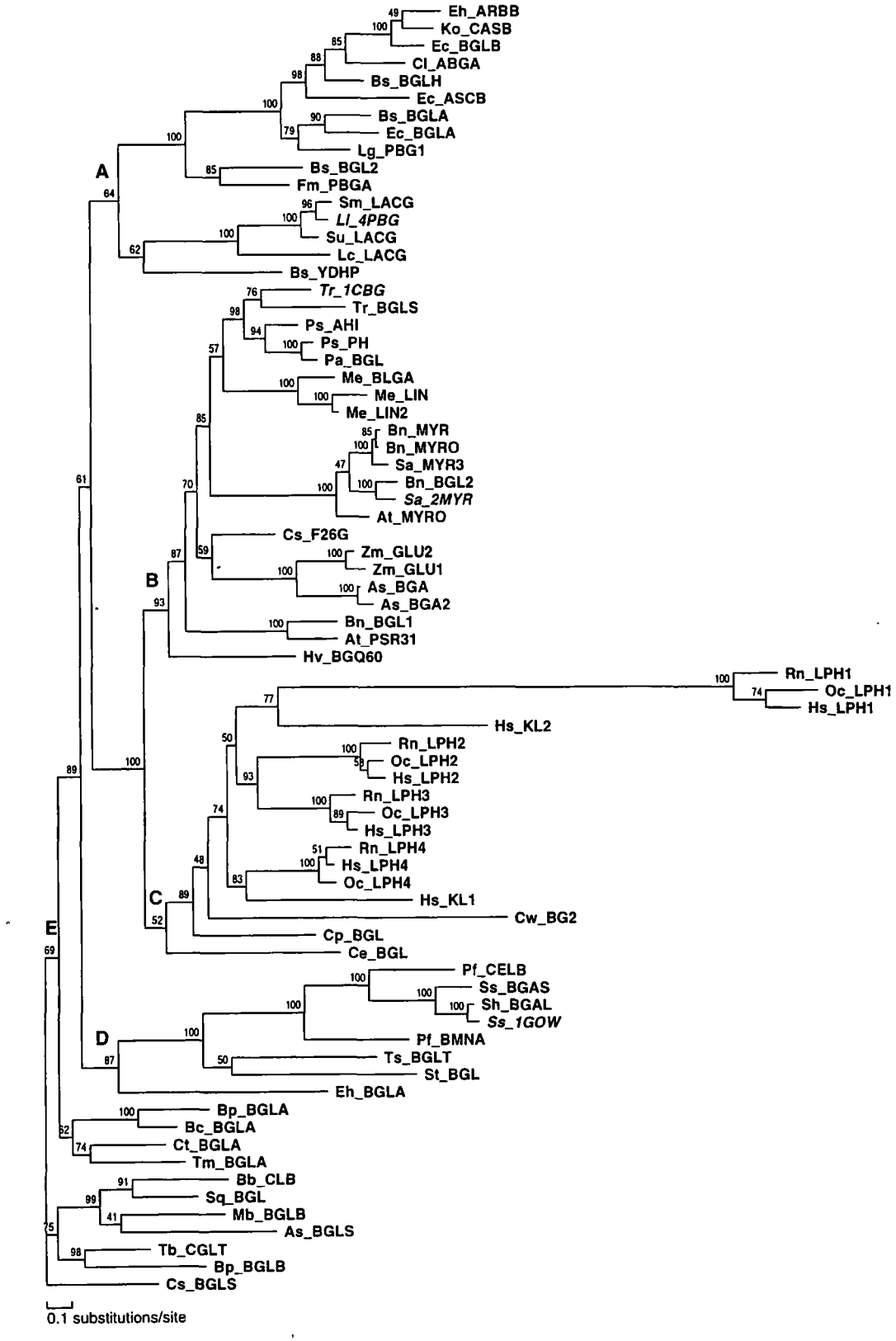


Figure 1. Phylogenetic tree for the family 1 glycosidase domain computed using the alignment shown in Figure 2. Sequence identifiers are given in Table 1 and those in italics are glycosidases whose three-dimensional structures have been determined. Subfamilies discussed in the text are labeled at their root. Local bootstrap probabilities are given for each branch and indicate the bootstrap probability of that branch.

356, highlighted). In spite of this low level of conservation and as is the case with other family 1 glycosidases, any reaction catalyzed by Klotho should retain stereochemistry.

The four glycosidases for which structures have been determined are 468-501 residues long. However, only 167 of the 356 positions in the domain (~35% of the total number of residues) form the core of the $(\beta/\alpha)_8$ barrel (red, labelled β1-β8, α1-α8 in Figure 2). Among the remainder are α-helices and β-strands (cyan) that form the core of the glycosidase domain (β1(α1), β1(α2), β2(α1), β3(α1), β3(α2), β4(α1) and β8(β1), β8(β2) respectively). Using only the 167 α-carbon atoms corresponding to the core barrel depicted in Figure 2, Sa_2MYR and Ss_1GOW can be superimposed with an RMSD of 0.8 Å. The results are shown in Figures 3, 4 and 5. Employing only residues in the core barrel for superimposition leads to a good spatial alignment of not only the core glycosidase β-strands and α-helices (cyan), but also the active site residues (side chains drawn explicitly).

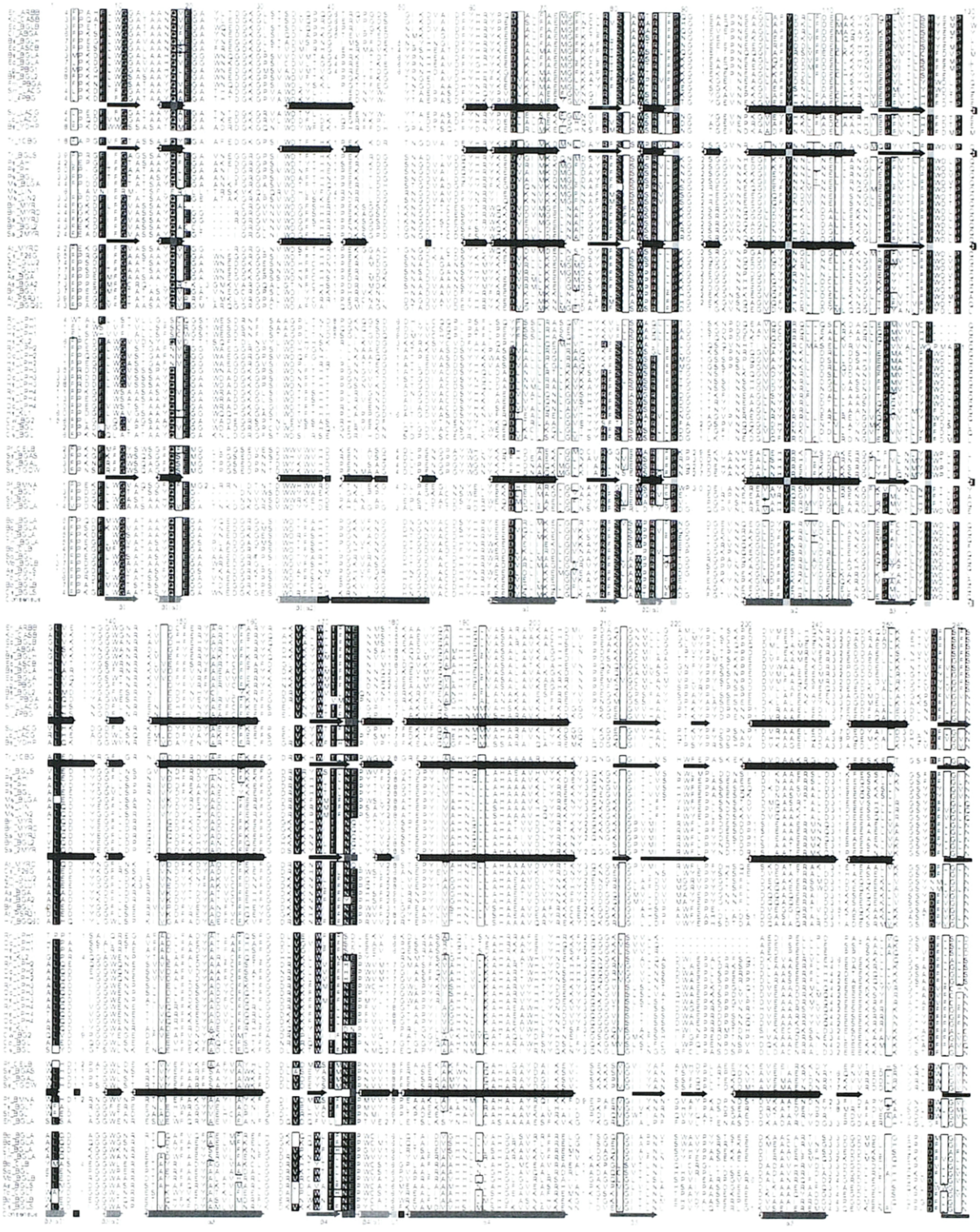
In both enzymes, the core glycosidase elements are located at the end of the barrel containing the active site. The tetramer interface in Ss_1GOW (7) and the dimer interface in Sa_2MYR (15) (blue) are located at similar positions on the outer surface of the domain. Since lactase-phlorizin hydrolase and Klotho are the only family members that possess two or more copies of the domain, the analogous regions may be the sites of interaction between the constituent domains. Thus, these enzymes may be a pseudotetramer and pseudodimer, respectively. It is possible that two Klotho molecules associate to form a dimer of pseudodimers that results in juxtaposition of the amino-termini extracellularly and carboxy-termini intracellularly. This association could be triggered by ligand binding and thus be an important mechanism underlying the biological activity of Klotho. With regards to the L1 and L2 loops (yellow), their variability at both the primary sequence and tertiary structure levels suggests that they could form flexible flaps over the the active site channel in all domains and thus play a key role in substrate recognition (particularly the non-reducing end).

The organization of and features observed in Ss_1GOW, Sa_2MYR and other family members provide a detailed framework for creating a three-dimensional model for the domains in Klotho and lactase-phlorizin hydrolase by homology modelling. The folding of these domains should create enzymes with similar, though not identical, active sites whose differences are likely to be important.

Active site. Figure 5 shows the great similarity in the geometry of active site residues in a eukaryotic *S*-glycosidase and an archaeal *O*-glycosidase. Side chains shown in magenta and green and the corresponding residues in other family members (Figure 2) have been discussed previously because of their roles in substrate binding and hydrolysis (1-3, 7, 10, 14, 15). Together with loops L1 and L2, differences between these residues are likely to play a role in precise substrate recognition. The nucleophile of family 1 glycosidases is a Glu at the end of β7. Inspection of Figure 2 indicates that some members of subfamily C differ at this position. The Asp in the first domain of lactase-phlorizin hydrolases and the serine (Ser) in the second domain of Klotho (Hs_KL2) are residues that could behave as nucleophiles. In the second domain of lactase-phlorizin hydrolase, however, unless some post-transcriptional event alters the glycine (Gly) at this position, this domain may be unable to perform the first glycosylation step, *i.e.*, it may be able to bind but not cleave its substrate. The Glu/Gln between β4 and β4(α1) implicated in positioning the nucleophilic water molecule (15) is changed to Asp in the first domain of lactase-phlorizin hydrolases and asparagine (Asn) in Hs_KL1. This observation supports the proposition that hydrolysis of the glycosyl-enzyme intermediate of retaining glycosidases may not necessarily involve base activation of the water molecule.

In contrast to what might be expected, the four invariant and most highly conserved positions (yellow) do not include the proton donor and nucleophile implicated in catalysis. These tyrosine (Tyr), histidine (His), proline (Pro) and aspartate

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174



(Asp) residues have not been examined to date and are thus good candidates for site-directed mutagenesis studies aimed at elucidating their roles in the structure, function and folding of the domain. The proximity of the His to the scissile bond suggests that rather than being involved in substrate recognition as in Sa_2MYR (15), it may be required for stabilization of the transition rather than ground

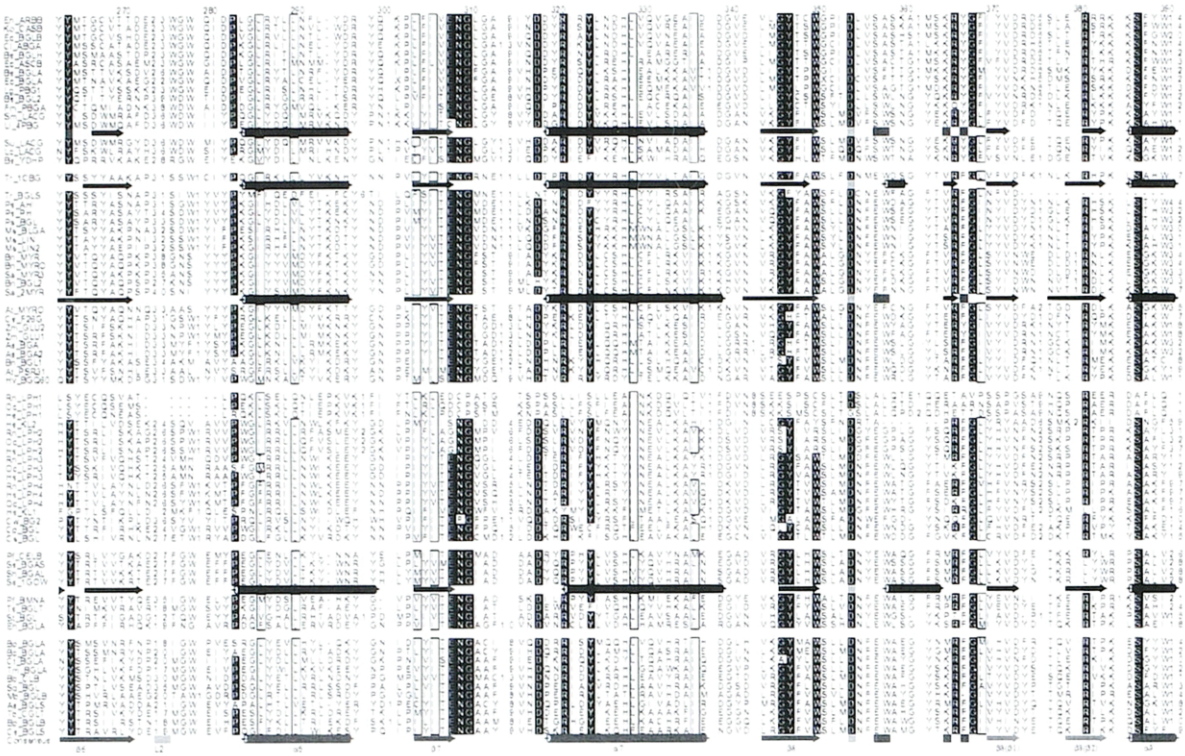
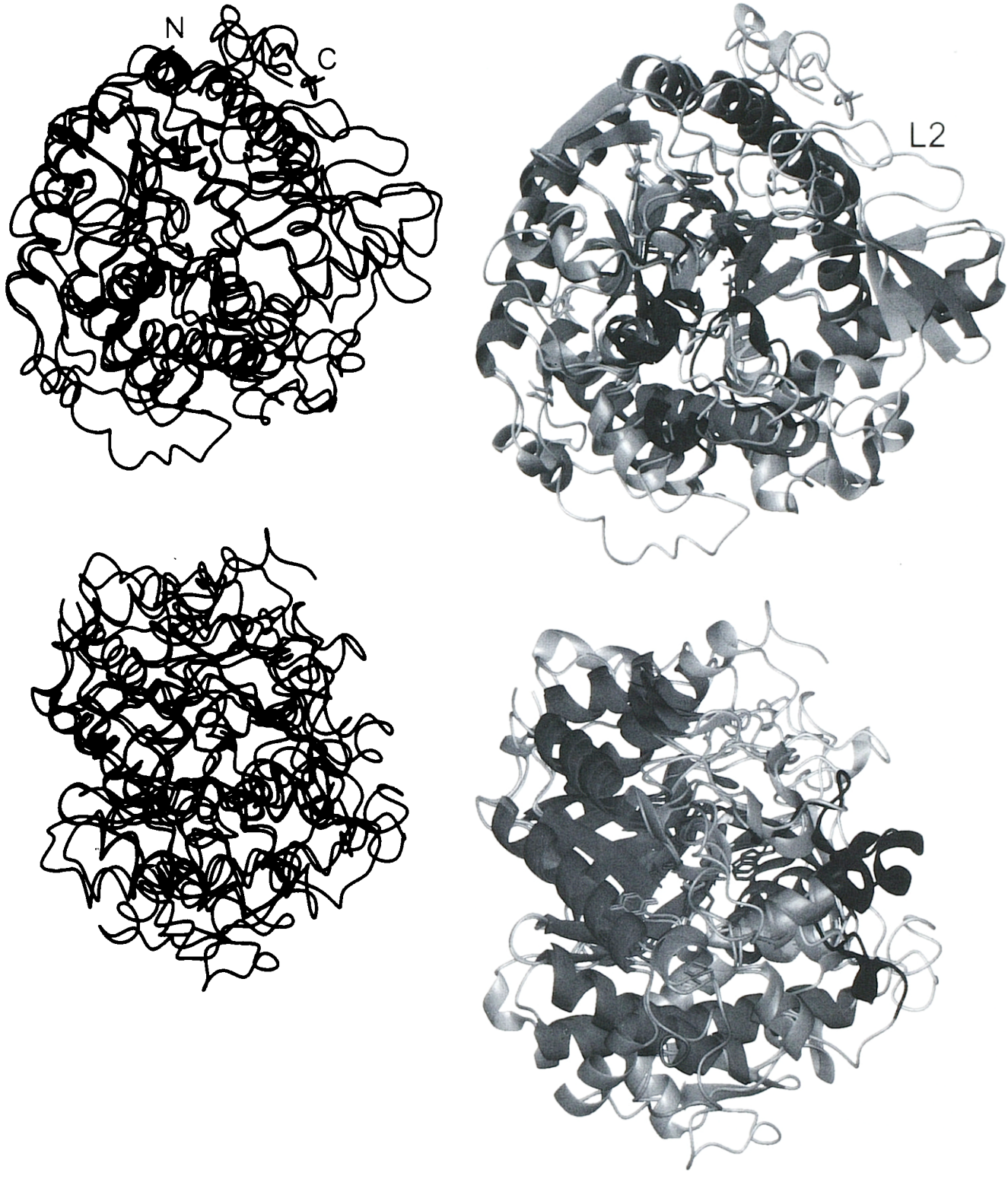


Figure 2. An HMM-generated alignment of the family 1 glycosidase domains listed in Table 1. Amino acids conserved in the majority of the sequences are highlighted; those in yellow are the invariant or most highly conserved residues. Positions that are predominantly hydrophobic are boxed. Columns containing "." correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present). For the four family members whose structures have been determined by X-ray crystallography, the locations of β-strands (arrows) and α-helices (cylinders), taken from the CATH database (67), are shown below each sequence (as given in the PDB entry). The structures are as follows. L1_4PBG: *Lactococcus lactis* 6-phospho-β-galactosidase and a bound galactose-6-phosphate molecule (14). The substrate was used as the basis for modeling the binding mode of the educt (lactose-6-phosphate) (14). Note that the glutamic acid nucleophile at the end of β7 was mutated to cysteine. Tr_1CBG: *Trifolium repens* cyanogenic β-glucosidase (10). Sa_2MYR: *Sinapis alba* myrosinase and a covalent 2-deoxy-2-fluorglucosyl-enzyme intermediate (15). Ss_1GOW: *Sulfolobus solfataricus* β-glycosidase (7). Positions that interact with the glucose substrate are magenta and the active site nucleophile and proton donor are green. Residues in blue are zinc ion ligands at the dimer interface in Sa_2MYR and at the tetramer interface in Ss_1GOW. The consensus structure for this glycosidase domain as deduced from the four known structures is shown. Secondary structure elements forming the core $(\beta/\alpha)_8$ barrel are shown in red and labeled β1-β8 and α1-α8. Strands and helices that are not part of the core barrel but are part of the core glycosidase domain are in cyan. In the core glycosidase domain, L1 and L2 (yellow) are two loop regions of variable length proposed to be involved in substrate recognition and binding, magenta positions are predicted to be involved primarily in recognition of the reducing end of the substrate, green positions are the nucleophile and proton donor and blue positions are those likely to be present at a subunit interface. Blank lines demarcate the subfamilies identified by phylogenetic analysis (the domains are ordered according to the tree shown in Figure 1). The numbers at the top mark every tenth column in this figure and not every tenth node in the HMM.

state. A similar function has been ascribed to three His residues in the active site of cyclodextrin glucanotransferase which catalyzes the formation of cyclodextrins from amylose through an intramolecular transglycosylation reaction (50). An alternative and/or additional role for the His may be positioning of the water molecule required for deglycosylation. The three other conserved residues are far removed from the active site and thus unlikely to have a direct role in catalysis. The invariant Pro is likely to be important for the structure and folding of the enzyme. The functions of the Asp and invariant Tyr remain to be determined. These results highlight the importance of

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174



Figures 3 and 4

Figure 5. Stereoviews of the structurally and/or functionally important side chains in the two family 1 glycosidases shown in Figures 3 and 4. The residues are colored according to Figure 2 and are the same as those drawn explicitly in Figure 4. The 2-deoxy-2-fluorglucosyl substrate bound to myrosinase is shown with carbon atoms in grey, oxygen in red and fluorine in green.

Figure 3. Orthogonal views of the overall structures of *Sinapis alba* myrosinase (red, Sa_2MYR) and *Sulfolobus solfataricus* β-glycosidase (blue, Ss_1GOW). The locations of the amino- and carboxy-terminii in both structures are labeled N and C respectively. The enzymes were superimposed using only the 167 α-carbon atoms corresponding to the core barrel depicted in Figure 2 (the secondary structure elements shown in red and labeled β1-β8 and α1-α8). The axis of the $(\beta/\alpha)_8$ barrel is perpendicular (top) and parallel (bottom) to the plane of the page.

Figure 4. The similarities and differences between the two family 1 glycosidases shown in Figure 3 (the enzymes are in the same orientations). The coloring scheme is that given in Figure 2. Ribbons correspond to the β-strands and α-helices of the barrel (red) and family 1 glycosidase domain (cyan) (those in grey are elements specific to each structure). Side chains drawn explicitly in magenta, yellow and green are important for structure and/or function. L2 (yellow) denotes a loop region proposed to be important in substrate recognition. The 2-deoxy-2-fluorglucosyl substrate bound at the active site of myrosinase is shown with carbon atoms in grey, oxygen in red and fluorine in green.

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174

characterizing both the variable and conserved regions of a family and the necessity for examining the entire domain rather than sequence motifs.

Family 30/59 Glycosidase Domain: Gaucher and Krabbe Disease Proteins

Hidden Markov model and secondary structure prediction. PSI-BLAST (36) searches using the Gaucher disease protein as the query sequence revealed statistically significant similarities to some bacterial and worm glycosidases and Krabbe disease protein. These results suggested a common underlying architecture among these proteins so these sequences were used to train a family 30/59 glycosidase domain HMM. The final HMM had a total of 351 nodes, approximately the length of the family 1 glycosidase domain HMM. Since no structural data are available for any members of family 30/59, residues corresponding to match and delete states in an HMM-generated alignment of the 14 training set sequences were used to predict the secondary structure for this domain using two different methods: DSC (51) and PHD (52). These predictions were combined to create a consensus secondary structure which indicated an alternating pattern of β-strands and α-helices followed by a series of β-strands. Since many glycosidases appear to have a $(\beta/\alpha)_8$ barrel fold, this alternating pattern could indicate a similar fold. The locations of the nucleophile and proton donor at the ends of β4 and β7 in family 1 (Figure 2) were employed as constraints on the placement of specific β-strands and α-helices in a barrel. The consensus secondary structure and patterns of residues from family 1 were used to infer/estimate the core of the $(\beta/\alpha)_8$ barrel. The results are shown in Figure 6.

Fold prediction. A further set of experiments were performed to assess whether the domain could have a $(\beta/\alpha)_8$ barrel fold. For each glycosidase, a sequence consisting of residues corresponding to match states (those shown in Figure 6) were employed as input to 123D (53). 123D determines a plausible fold for a protein of unknown structure from a library of representative protein structures. It uses a substitution matrix, secondary structure prediction and contact capacity potentials to thread a sequence through a set of structures. For each sequence, although few structures from the library had Z-scores significantly greater than 0.0, among the top 10 structures were one or more glycosidases that have been classified (1) as follows:

- Family 5: endoglucanases and β-mannanases *Clostridium thermocellum* endo-1,4-β-glucanase C307 (cellulase) (PDB entry 1CEO; Swiss-Prot entry GUNC_CLOSF).
- Family 6: endoglucanases and cellobiohydrolases *Thermomonospora fusca* strain YX endo-1,4-β-D-glucanase E-2 (1TML; GUN2_THEFU).
- Family 10: mostly xylanases *Clostridium thermocellum* 1,4-β-D-xylan-xylanohydrolase Z (1XYZ; XYNZ_CLOTM). *Pseudomonas fluorescens* 1,4-β-D-xylan xylanohydrolase A (1CLX; XYNA_PSEFL).
- Family 13: α-amylases, pullulanases, cyclomaltodextrin glucanotransferase, cyclomaltodextrinase and trehalose-6-phosphate hydrolase *Pseudomonas saccharophila* glucan 1,4-α-maltotetrahydrolase (2AMG; AMT4_PSESA). *Hordeum vulgare* 1,4-α-D-glucan glucanohydrolase (1AMY; AMY2_HORVU).
- Family 14: β-amylases *Glycine max* 1,4-α-D-glucan maltohydrolase (1BYB; AMYB_SOYBN).
- Family 18: chitinases, endo-β-N-acetylglucosaminidases *Flavobacterium meningosepticum* endo-β-N-acetylglucosaminidase F1 (2EBN; EBA1_FLAME).
- Unknown family *Flavobacterium meningosepticum* peptide-N(4)-(N-acetyl-β-D-glucosaminyl) asparagine amidase F (1PGS; PNGF_FLAME).

Apart from 1PGS, the SCOP database of structures (54), categorizes these structures as possessing a β/α barrel fold or its close variant the cellulase fold.

Taken together, the data support the proposal that the family 30/59 glycosidase domain has a $(\beta/\alpha)_8$ barrel fold similar to that of the family 1 glycosidase domain. Differences between the two folds arise from variations in the lengths of the β−α connections.

Disease mutations and active site. The active site and substrate binding regions in the family 1

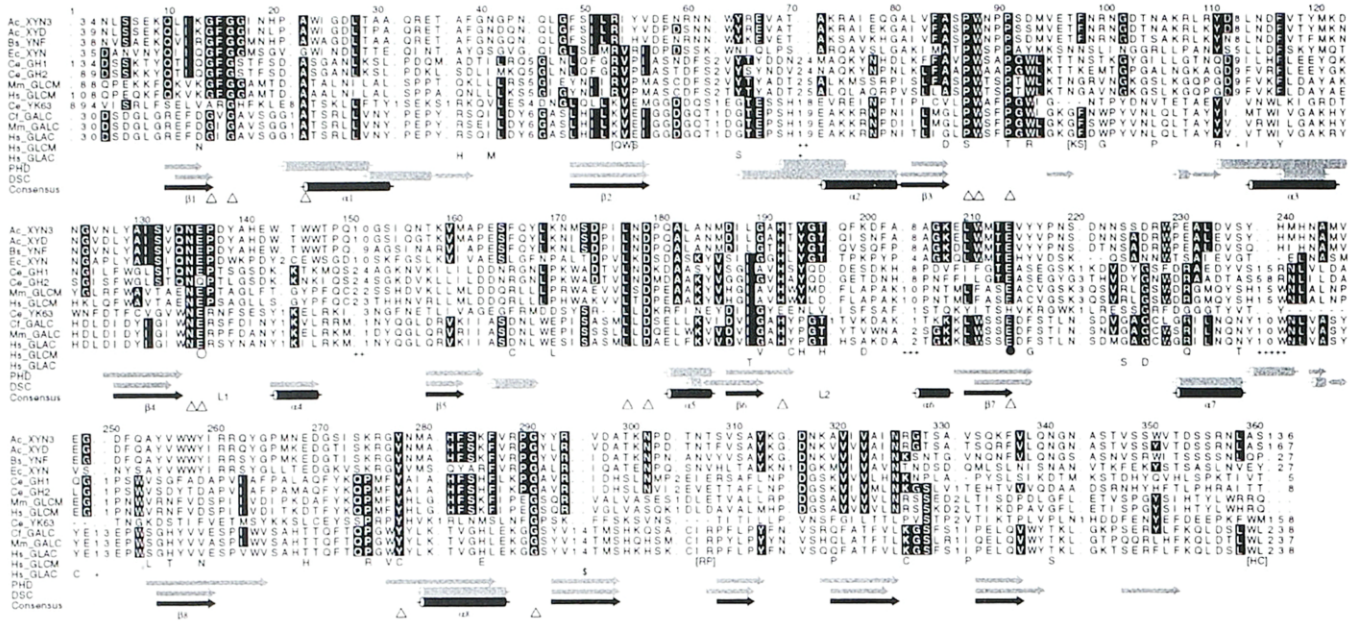


Figure 6. An HMM-generated alignment of the family 30/59 glycosidase domain. Invariant or conserved residues are highlighted. The sequences shown are as follows (databank codes are in square parenthesis). Ac_XYN3, *Aeromonas caviae* endo-xylanase [D88553]. Ac_XYD, *A. caviae* xylanase D [ACU86340]. Bs_YNF, *Bacillus subtilis* ORF YnfF [BC170DEGR]. Ec_XYN, *Erwinia chrysanthemi* xylanase [ECU41750]. Ce_GH1 and Ce_GH2, *Caenorhabditis elegans* glycosidases (ORFs C33C12.8 and C33C12.3) [CELC33C12]. Mm_GLCM, *Mus musculus* glucosylceramidase precursor [GLCM_MOUSE]. Hs_GLCM, *Homo sapiens* glucosylceramidase precursor (Gaucher disease) [GLCM_HUMAN]. Ce_YK63, *C. elegans* galactosylceramidase-like ORF [YK63_CAEEL]. Cf_GALC, *Canis familiaris* galactosylceramidase precursor [GALC-CANFA]. Mm_GALC, *M. musculus* galactosylceramidase precursor [GALC_MOUSE]. Hs_GLAC, *H. sapiens* galactosylceramidase precursor (Krabbe disease) [GALC_HUMAN]. Not shown are galactosylceramidase from *Macaca mulatta* [MMGALC01] and a *C. elegans* ORF [CEF11E6] that appears to be the amino-terminal region of a glucosylceramidase. Mm_GLCM and Hs_GLCM belong to family 30 while Cf_GALC, Mm_GALC and Hs_GLAC belong to family 59. Disease mutations occurring in Hs_GLCM and Hs_GLAC are taken from their respective databank files and elsewhere (68, 69). Amino acids in parentheses denote the altered residues at a given site, $ indicates a termination codon and each + marks a mutation occurring at a position that is part of an insertion in the HMM. The open and filled circles indicate the predicted proton donor and nucleophile respectively and triangles indicate the conserved or invariant positions. The α-helices (cylinders) and β-strands (arrows) predicted by DSC (51) and PHD (52) are shown. Consensus indicates the secondary structure inferred for this glycosidase domain. Secondary structure elements predicted to form the core of the $(\beta/\alpha)_8$ barrel are labeled β1-β8 and α1-α8. L1 and L2 denote the regions that would correspond to the putative substrate binding loops in the family 1 glycosidase domain (Figure 2).

glycosidase domain are located at the C-terminal end of the barrel (Figure 4), suggesting a similar arrangement in the family 30/59 glycosidase domain. Given the proximity of L2 to the C-terminus in the family 1 glycosidase domain, it is possible that the β-strands following the barrel in family 30/59 may participate in substrate recognition and binding. A mutation that creates a termination codon in Hs_GLAC would generate a protein corresponding to the domain proposed here and thus enzymatic activity would remain unaffected ($ in Figure 6). The β−α connections forming the active site and substrate binding regions contain the majority of mutations in both Gaucher and Krabbe diseases and/or the conserved and invariant positions (marked by open triangles). The latter include Pro, Tyr, Asp, Trp and His residues likely to be important in substrate binding and or catalysis and thus the functional equivalents of the corresponding conserved residues in family 1 (see Figures 2 and 5). The disease mutations are more likely to affect the function of the enzyme rather than the structural integrity of the barrel.

Blood Cells, Molecules, and Diseases (1998) 24(7) Apr 15: 83–100
Article No. MD980174

DISCUSSION

The family 1 glycosidase domain HMM trained here attempts to capture the core elements of the $(\beta/\alpha)_8$ barrel global fold and the residues involved in recognizing the carbohydrate residue adjacent to the scissile bond. While knowledge of these common features is a necessary early step in providing greater insights into this family, particularly members such as Klotho and lactase-phlorizin hydrolase, it is insufficient for a comprehensive understanding. Clearly, the precise substrate specificity and mode of action of each member will be governed to a large degree by the more variable parts of the structure, most notably the L1 and L2 loops. The flexibility of HMMs means that it is simple and straightforward to model these variable regions explicitly (in more detail) and generate alignments for further examination. Alternatively, if the goal is detecting more remote homologues and or characterizing the $(\beta/\alpha)_8$ barrel, the connecting regions within and between the β/α repeats could be modeled implicitly by converting them to insertions leaving only the β-strands and α-helices of the core barrel. The latter strategy would be most suitable for identifying distant relationships by merging specific glycosidase families and superfamilies in an effort to approximate the "archetypal" (or ancestral) glycosidase fold. In either analysis, extension of the current HMM generates automatically a statistical model that can be used for database searching. HMMs provide a framework for comparative analysis of genomes/proteins as well as detailed analysis of specific proteins. A synthesis of the considerable information on the family 1 domain and data from studies of the family 30/59 glycosidase domain demonstrate the ability of computational studies to provide insights into Gaucher and Krabbe diseases.

The results here provide some clues into cellular senescence and thus indirectly into the complex, multifactorial phenomenon of organismal aging. In addition to Klotho, some of the eucaryotic family 1 glycosidases in subfamily C (Figure 1) also appear to be associated with aging. For example, defects in human lactase-phlorizin hydrolase are the cause of disaccharide intolerance type II or III and the activity of this enzyme sometimes declines in adults (55). There is an apparent age-related decline in its activity in rats (56). One of the plant enzymes (Zm_GLU1) cleaves the biologically inactive plant hormone conjugates cytokinin-*O*-glucosides and kinetin-N3-glucoside, releasing the active cytokinins (57). Cytokinins, compounds structurally related to adenine, are found in many plants, in bacteria and in the tRNA of many bacteria and eukaryotes. In plants, they appear to promote cell division and differentiation and are associated with delay of senescence and the promotion of chloroplast and lateral bud development (reviewed in (58-60)). The cytokinin kinetin (N6-furfuryladenine) retards senescence in plants, delays aging in human cells in culture, slows development of insects and prolongs their lifespan (61-63). Recently, kinetin has been detected in commercially available DNA, human cellular DNA and plant cell extracts (64). A mechanism for the *in vivo* formation of kinetin has been suggested: it is a secondary oxdiative damage product of DNA (65, 66).

The known or inferred glycosidase activities of Klotho, lactase-phlorizin hydrolase, Gaucher disease protein and Krabbe disease protein include hydrolysis of membrane phospholipids thereby generating ceramide, a second messenger that activates the apoptotic cascade. Whether these four mammalian enzymes also have a role in the generation of kinetin is unknown. The data here provide a foundation for modeling the three-dimensional structures of the glycosidase domains present in, among others, these enzymes. Although these and other computational and experimental studies can be useful in understanding the normal and abnormal phenotypes at the molecular level, it will be a challenge translating these observations into an understanding of the cellular, organismal and clinical aspects of aging, cancer and certain human disorders.

ACKNOWLEDGMENTS

I thank G.M. Martin for bringing Klotho to my attention. This work was supported by the Director,

Office of Energy Research, Office of Biological and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76F00098 (ISM). The data and multiple alignments are available in electronic form upon request.

REFERENCES

1. Henrissat B, and Bairoch A. Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* 316:695-696, 1996. [An index of glycosyl hydrolase entries in SWISS-PROT can be obtained at the following URL http://www.expasy.ch/cgibin/lists?glycosid.txt.
2. Sinnott M. Catalytic mechanisms of enzymatic glycosyl transfer. *Chem Rev* 90:149-154, 1990.
3. Davies G, Henrissat B. Structures and mechanisms of glycosyl hydrolases. *Structure* 3:853-859, 1995.
4. McKusick V. Online Mendelian Inheritance in Man, OMIM. Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). 1997. [The OMIM database is available at URL http://www3.ncbi.nlm.nih.gov/omim/.]
5. Kuro-o M, Matsumura Y, Aizawa H, Kawaguchi H, Suga T, Utsugi T, Ohyama Y, Kurabayashi M, Kamane T, Kume E, Iwasaki H, Iida A, Shiraki-Iida T, Nishikawa S, Nagai R, and Nabeshima Y. Mutation of the mouse *klotho* gene leads to a syndrome resembling ageing. *Nature* 390:45-51, 1997.
6. Martin G, and Mian I. Ageing: New mice for old questions. *Nature* 390:18-19, 1997.
7. Aguilar C, Sanderson I, Moracci M, Ciaramella M, Nucci R, Rossi M, Pearl L. Crystal structure of the β-glycosidase from the hyperthermophilic archeon *Sulfolobus solfataricus:* Resilience as a key factor in thermostability. *J Mol Biol* 271:789-802, 1997.
8. Kengen S, Starms A. An extremely thermostable β-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus:* A comparison with other glycosidases. *Biocatalysis* 11:79-88, 1994.
9. Gabelsberger J, Liebl L, Schleifer K. Purification and properties of recombinant β-glucosidase of the hyperthermophilic bacterium *Thermotoga maritima. Appl Microbiol Biotechnol* 40:44-52, 1993.
10. Barrett T, Suresh C, Tolley S, Dodson E, Hughes M. The crystal structure of a cyanogenic β-glucosidase from white clover, a family 1 glycosyl hydrolase. *Structure* 3:951-960, 1995.
11. Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon J, Davies G. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc Natl Acad Sci* 92:7090-7094, 1995. [Published erratum appears in 93(11): 5674 (1996)]
12. Jenkins J, Lo Leggio L, Harris G, Pickersgill R. β-glucosidase, β-galactosidase, family A cellulases, family F xylanases and two barley glycanases form a superfamily of enzymes with 8-fold β/α architecture and with two conserved glutamates near the carboxy-terminal ends of β-strands four and seven. *Febs Lett* 362:281-285, 1995.
13. Gerloff D, Benner S. A consensus prediction of the secondary structure for the 6-phospho-β-D-galactosidase superfamily. *Proteins* 21:273-281, 1995.
14. Wiesmann C, Hengstenberg W, Schulz G. Crystal structures and mechanism of 6-phospho-β-galactosidase from *Lactococcus lactis. J Mol Biol* 269:851-860, 1997.
15. Burmeister W, Cottaz S, Driguez H, Iorie R, Palmieri S, Henrissat B. The crystal structures of *Sinapis alba* myrosinase and a covalent glycosyl-enzyme intermediate provide insights into the substrate recognition and active-site machinery of an *S*-glycosidase. *Structure* 5:663-675, 1997.
16. Banner D, Bloomer A, Petsko G, Phillips D, Pogson C, Wilson I, Corran P, Furth A, Milman J, Offord R, Priddle J, Waley S. Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5Å resolution using amino acid sequence data. *Nature* 255:609-614, 1975.
17. Krogh A, Brown M, Mian I, Sjölander K, Haussler D. Hidden Markov models in computational biology: Applications to protein modelling. *J Mol Biol* 235:1501-1531, 1994.
18. Hughey R, Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12:95-107, 1996. [The hidden Markov model software can be accessed at URL http://www.cse.ucsc.edu/research/compbio/sam.html.]
19. Brown M, Hughey R, Krogh A, Mian I, Sjölander K, Haussler D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *ISMB*, 1:47-55, 1993.
20. Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian I, Haussler D. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *CABIOS* 12:327-345, 1996.
21. Mian I. Sequence analysis of ribonucleases HII, III, II, PH and D. *Nucleic Acids Res* 25:3187-3195, 1997.
22. Moser M, Holley W, Chatterjee A, Mian I. The proofreading domain of *Escherichia coli* DNA polymerase I and other DNA and/or RNA exonuclease domains. *Nucleic Acids Res* 25:5110-5118, 1997.
23. Mian I, Moser M, Holley W, Chatterjee A. Statistical modelling and phylogenetic analysis of a deaminase domain. *J Comp Bio* 5:59-75, 1998.

24. Mian I, Moser M. The Fanconi anaemia complementation group A protein contains a peroxidase domain. *Mol Genet Metabol* 1998 [in press]
25. Dalgaard J, Moser M, Hughey R, Mian I. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J Comp Bio* 4:193-214, 1997.
26. Dalgaard J, Klar A, Moser M, Holley W, Chatterjee A, Mian I. Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res* 25:4626-4638, 1997.
27. Herbert A, Alfken J, Kim Y-G, Mian I, Nishijura K, Rich A. A Z-DNA binding domain present in human editing enzyme, double-stranded RNA adenosine deaminase. *Proc Natl Acad Sci* 94:8421-8426, 1997.
28. Bateman A, Chothia C. Fibronectin type III domains in yeast detected by a hidden Markov model. *Curr Biol* 6:1544-1547, 1996.
29. Bateman A, Eddy S, Chothia C. Members of the immunoglobulin superfamily in bacteria. *Protein Sci* 5:1939-1941, 1996.
30. Hazes B. The (QxW)3 domain: A flexible lectin scaffold. *Protein Sci* 5:1490-1501, 1996.
31. Shub D, Goodrich-Blair H, Eddy S. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *TIBS* 19:402-404, 1994.
32. Grundy W, Bailey T, Elkan C, Baker M. Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochem Biophys Res Commun* 231:760-766, 1997.
33. Baldi P, Chauvin Y, Hunkapiller T, and McClure M. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci* 91:1059-1063, 1994.
34. Eddy S. Multiple alignment using hidden Markov models *ISMB* 3:114-120, 1995. [The hidden Markov model software can be accessed at URL http://genome.wustl.edu/eddy/hmm.html.]
35. Fujiwara Y, Asogawa M, Konagaya A. Stochastic motif extraction using hidden Markov model. *ISMB* 2:121-129, 1994.
36. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402, 1997.
37. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol* 215:403-410, 1990.
38. Altschul S. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219:555-565, 1991.
39. Barrett C, Hughey R, Karplus K. Scoring hidden Markov models. *CABIOS* 13:191-199, 1997.
40. NCI NRP (Non-Redundant Protein) and NRN (Non-Redundant Nucleic Acid) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute's Frederick Biomedical Supercomputing Center, 1998.
41. Adachi J. Modelling of molecular evolution and maximum likelihood inference of molecular phylogeny PhD dissertation Institute of Statistical Mathematics, Tokyo, 1995.
42. Adachi J, Hasegawa M. MOLPHY: Programs for Molecular Phylogenetics, I. PROTML: Maximum Likelihood Inference of Protein Phylogeny Computer Science Monographs 27, Institute of Statistical Mathematics, Tokyo, 1992. [MOLPHY is available from ftp://sunmh.ism.ac.jp/pub/molphy.]
43. Barton G. ALSCRIPT—A tool to format multiple sequence alignments. *Protein Engineer* 6:37-40, 1993.
44. Maciukenas M. Treetool: An interactive tool for displaying, editing and printing phylogenetic trees, 1992. [Currently, Treetool is modified and maintained by Mike McCaughey, Ribosomal Database Project, University of Illinois. It is available from ftp://rdp.life.uiuc.edu/rdp/programs/TreeTool.]
45. Koradi R, Billeter M, Wuthrich K. MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graphics* 14:51-55, 1996.
46. Mantei N, Villa M, Enzler T, Wacker H, Boll W, James P, Hunziker W, Semenza G. Complete primary structure of human and rabbit lactase-phlorizin hydrolase: Implications for biosynthesis, membrane anchoring and evolution of the enzyme. *EMBO J* 7:2705-2713, 1988.
47. Skory C, Freer S. Cloning and characterization of a gene encoding a cell-bound, extracellular β-glucosidase in the yeast *Candida wickerhamii. Appl Environ Microbiol* 61:518-525, 1995.
48. Skory C, Freer S, Bothast R. Expression and secretion of the *Candida wickerhamii* extracellular β-glucosidase gene, *bglB,* in *Saccharomyces cerevisiae. Current Genetics* 30:417-422, 1996.
49. Hays W, Jenison S, Yamada T, Pastuszyn A, Glew R. Primary structure of the cytosolic β-glucosidase of guinea pig liver. *Biochem J* 319:829-837, 1996.
50. Nakamura A, Haga K, Yamane K. Three histidine residues in the active center of cyclodextrin glucanotransferase from alkalophilic *Bacillus* sp. 1011: Effects of the replacement on pH dependence and transition-state stabilization. *Biochemistry* 32:6624-631, 1993.
51. King R, Sternberg M. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* 5:2298-2310, 1996.

52. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55-72, 1994.
53. Alexandrov N, Nussinov R, Zimmer R. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials In L. Hunter and T.E. Klein, (ed.), Pacific Symposium on Biocomputing '96, pp. 53-72 World Scientific Publishing Company Singapore, 1995. [The 123D program is available at URL http://www-lmmb.ncifcrf.gov/~nicka/123D.html.]
54. Murzin A, Brenner S, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536-540, 1995.
55. Sahi T. Hypolactasia and lactase persistence: Historical review and the terminology. *Scand J Gastroenterol* 202:1-6, 1994.
56. Lee M,.Russell R, Montgomery R, Krasinski S. Total intestinal lactase and sucrase activities are reduced in aged rats. *J Nutr* 127:1382-1387, 1997.
57. Brzobohaty B, Moore I, Kristoffersen P, Bako L, Campos N, Schell J, Palme K. Release of active cytokinin by a β-glucosidase localized to the maize root meristem. *Science* 262:1051-1054, 1993.
58. Brzobohaty B, Moore I, Palme K. Cytokinin metabolism: implications for regulation of plant growth and development. *Plant Mol Biol* 26:1483-1497, 1994.
59. Schell J, Koncz C, Spena A, Palme K, Walden R. Genes involved in the control of growth and differentiation in plants. *Gene* 135:245-249, 1993.
60. Costacurta A, Vanderleyden J. Synthesis of phytohormones by plant-associated bacteria. *Crit Rev Microbiol* 21:1-18, 1995.
61. Rattan S, Clark B. Kinetin delays the onset of ageing characteristics in human fibroblasts. *Biochem Biophys Res Commun* 201:665-672, 1994.
62. Sharma S, Kaur P, Rattan S. Plant growth hormone kinetin delays ageing, prolongs the lifespan and slows down development of the fruitfly *Zaprionus paravittiger*. *Biochem Biophys Res Commun* 216:1067-1071, 1995.
63. Sharma S, Kaur J, Rattan S. Increased longevity of kinetin-fed *Zaprionus* fruitflies is accompanied by their reduced fecundity and enhanced catalase activity. *Biochem Mol Biol Int* 41:869-875, 1997.
64. Barciszewski J, Siboska G, Pedersen B, Clark B, Rattan S. Evidence for the presence of kinetin in DNA and cell extracts. *Febs Lett* 393:197-200, 1996.
65. Barciszewski J, Siboska G, Pedersen B, Clark B, Rattan S. A mechanism for the *in vivo* formation of N6-furfuryladenine, kinetin, as a secondary oxidative damage product of DNA. *Febs Lett* 414:457-460, 1997.
66. Barciszewski J, Siboska G, Pedersen B, Clark B, Rattan S. Furfural, a precursor of the cytokinin hormone kinetin, and base propenals are formed by hydroxyl radical damage of DNA. *Biochem Biophys Res Commun* 238:317-319, 1997.
67. Orengo C, Michie A, Jones S, Swindells M, Jones D, Thornton J. Cath: Protein structure classification, 1997. [The CATH database is available at URL http://www.biochem.ucl.ac.uk/bsm/cath]
68. Furuya H, Kukita Y, Nagano S, Sakai Y, Yamashita Y, Fukuyama H, Inatomi Y, Saito Y, Koike R, Tsuji S. Adult onset globoid cell leukodystrophy (Krabbe disease): Analysis of galactosylceramidase cDNA from four Japanese patients. *Human Genet* 100:450-456, 1997.
69. De Gasperi R, Gama Sosa M, Sartorato E, Battistini S, MacFarlane H, Gusella J, Krivit W, Kolodny E. Molecular heterogeneity of late-onset forms of globoid-cell leukodystrophy. *Am J Human Genet* 59:1233-1242, 1996.