Theor Appl Genet (1992) 84:643–655

THEORETICAL AND APPLIED GENETICS

Relationships among amino acid sequences of animal, microbial and plant peroxidases

H. Tyson

Biology Department, McGill University, Stewart Biology Building, 1205 Ave. Dr. Penfield, Montreal, Quebec, Canada, H3A 1B1

Received November 28, 1991; Accepted January 10, 1992
Communicated by K. Tsunewaki

Summary. Relationships among 18 peroxidases amino acid sequences of animal, microbial and plant origin were examined using optimum alignment of all pairwise sequence combinations to generate a total distance matrix. The matrix was used to cluster the sequences with complete linkage (farthest neighbour) procedures. Specific distances were calculated from the total distances matrix. The patterns of specific distances for each sequence were compared to evaluate the relationships between sequences, check their significance and construct subgroups of related sequences. The results were compared with those from clustering and its resultant dendrogram; good agreement was achieved. The 18 sequences fell into two principal groups, plant peroxidases and animal/microbial peroxidases. Within the plant peroxidases four subgroups were detected; the animal/microbial peroxidases formed a fifth subgroup. Profiles were constructed for the subgroups from lists of matching amino acids generated by the alignment calculations. Superimposed lists were realigned to recognise conserved areas and elements. Individual subgroup profiles for the plant peroxidases were then combined into a single profile which in turn was combined with profiles from the animal/microbial peroxidases. The final profile suggested that numerous sequence features (motifs) were common to peroxidases of widely different function and origins.

Key words: Peroxidases – Sequence relationships – Peroxidase profiles

Introduction

Peroxidases play a variety of important roles in animal, microbial and plant physiology, and continue to arouse interest in the molecular and structural characteristics that generate such versatility. They have a primary function to oxidise molecules at the expense of H_2O_2. Widely distributed in living organisms, numerous isoforms frequently exist. Besides peroxidatic oxidation of electron donor molecules, they can be involved in aerobic oxidations of dihydroxyfumarate, triose reductone and naphthohydroquinone, hydroxylation of aromatic molecules, formation of ethylene, halogenation and antimicrobial activity. Oxidation with H_2O_2, O_2 and polymerisation (condensation) have been documented. Plant peroxidases, in the cytoplasm and on membranes, break down H_2O_2 and oxidatively degrade the important plant hormone indole acetic acid (IAA) as well as forming cell-wall lignin by condensation of cinnamyl alcohols (see review by Gaspar et al. 1980). Ligninase depolymerases lignin through free radical formation (Harvey et al. 1985). A fungal chloride peroxidase catalyses the peroxidative halogenations involved in the biosynthesis of clardariomycin, acting as a peroxidase in the absence of chloride ions (Fang et al. 1986). Mammalian glutathione peroxidases, which have lengths of approximately 200 amino acids and a selenocysteine residue at the active site, protect haemoglobin in erythrocytes from oxidative breakdown, catalysing the reduction of H_2O_2 by glutathione (Gunzler et al. 1984). Human myeloperoxidase plays a major role in the oxygen-dependent microbicidal system of granulocytes; in the presence of H_2O_2 and chloride ions, it catalyses the production of hypochlorous acid. Human thyroid peroxidase operates as an oxidoreductase.

Most peroxidase sequences from animal, microbial and plant sources have lengths of 300–400 amino acids. Many peroxidases are glycoproteins with a median molecular mass in the 40,000–50,000 kDa range, and all contain a haem group located at the active site (except

the glutathione peroxidases, which have a selenocysteine residue at the active site). The range of functions displayed by peroxidases is largely dependent on the three-dimensional orientation of the haem group at the active site, as, for example, in the case of yeast cytochrome c peroxidase and fungal ligninase. Additional functional modifications are brought about by the number, location, size and composition of the oligosaccharide chains attached through asparagine linkages to the polypeptide core. These chains may be post-translationally modified, thus opening up further possibilities for functional fine tuning. They do not, however, appear to contribute to the correct folding of the protein (Smith et al. 1991).

Relationships among the large number of peroxidase (EC. 1.11.1.7) sequences now documented in databases and elsewhere were examined with numerical methods plus efforts to deduce the common features of their linear amino acid sequences. The aims here were (1) to determine peroxidase sequence relationships through calculation of total distances between sequences by optimum alignments in all pairwise combinations, (2) to cluster these total distances and construct a dendrogram from them, (3) to compute specific distances from total distances and evaluate the significance of sequence relationships and (4) to construct sequence profiles from the lists of matching amino acids supplied by the optimum alignments. In this way, it was hoped to see whether peroxidases fall into discernible subgroups of related sequences and whether there exist common features across all subgroups.

Eighteen peroxidase amino acid sequences were included in this study. The horseradish (Welinder 1979) and turnip (Mazza and Welinder 1980) sequences resulted from direct amino acid sequencing of the purified protein, but most amino acid sequences have been deduced from the corresponding DNA sequences. Peroxidase genes have been cloned and sequenced in numerous species, examples of which are listed below. In this list, peroxidases with functions differing from those occurring in plant species are specifically mentioned, for example glutathione peroxidase. The sequences (and their sources) compared in this study, representing a wide range of functions and sources, are as follows: *Arabidopsis thaliana* (Intrapruk et al. 1991), cattle (*Bos primigenius taurus*, glutathione peroxidase, Gunzler et al. 1984), *Caldariomyces fumago* (Chloride peroxidase, Fang et al. 1986), cucumber (*Cucumis sativus*, Morgens et al. 1990), man (*Homo sapiens*, myeloperoxidase, Morishita et al. 1987), horseradish (*Armoracia rusticana*, Fujiyama et al. 1988), peanut (*Arachis hypogaea*, Buffard et al. 1991), *Phanerochaete chrysosporium* (ligninase, Holzbaur and Tien, 1988; manganese peroxidase, Godfrey et al. 1990), potato (*Solanum tuberosum*; Roberts et al. 1988), pig (*Sus scrofa*, thyroid peroxidase, Magnusson et al. 1986), tobacco (*Nicotiana tabacum*, Lagrimini et al. 1987), tomato (*Lycopersicon esculentum*, Roberts and Kolattukudy 1989), yeast (*Saccharomyces cerevisae*, mitochondrial cytochrome c peroxidase, Kaput et al. 1982) and wheat (*Triticum aestivum*, Hertig et al. 1991).

Some of these peroxidases contain signal peptides that were removed before pairwise alignment. In some species, for example *Arabidopsis thaliana*, horseradish, peanut, tomato and wheat, isozymes of peroxidase have been sequenced. Preliminary investigation of the horseradish, tomato and wheat isozymes showed that isozyme sequences within these species are so similar that only one from each species need be taken as a representative of the peroxidase sequence for that species. The two peanut isozymes are markedly different, and in this case and that of the two *Arabidopsis thaliana* isozymes, both isozymes were included in the comparisons of the 18 sequences selected for this study. The 18 sequences, thus represented a sample to examine relationships and determine the essential sequence features of animal, microbial and plant peroxidases.

Optimum alignment techniques were used to make comparisons among the 18 peroxidases by calculating alignments for all pairwise sequence combinations. Simultaneous multiple optimum alignment for long sequences is not feasible, whereas procedures for pairwise optimum sequence alignment are thoroughly worked out (Sellers 1974). Each optimum alignment involves gap insertions, controlled by two user chosen gap penalties (weights) to limit excessive gap creation, but maximises matches. Penalties for (a) inserting a gap and (b) continuing the gap are set prior to alignments.

An optimum alignment supplies a total distance between two sequences. From a set of sequences and all their pairwise alignments, a square, symmetrical matrix of total distances is obtained. Conventionally, the total distance matrix is the input to sequence clustering techniques, generating a dendrogram or tree diagram to summarise relationships (Sneath and Sokal 1973). Dendrograms, branch points and branch distances were generated for these peroxidases by the appropriate software (SYSTAT 'cluster'; Wilkinson 1989).

An alternative approach draws on diallel cross analysis. Total distances from all pairwise alignments are analogous to progeny data from crosses in pairwise combinations (i.e. a diallel) among a set of parental genotypes, with parents and reciprocals absent from the F_1 generation. The total distances thus supply analogues to the general (GCA) and specific (SCA) combining abilities calculated from diallel crosses of the same format with Griffing's (1956) techniques. For total distances between sequences, equivalents to SCA values become specific distances, whose calculation has recently been described (Tyson 1991) and which are independent of the average of a sequence's distances. Specific distances generated by one sequence in all its interactions with the

other sequences in the set may then be compared with corresponding specific distances for any other sequence. Similar sequences generate similar specific distances; similarity in the numerical pattern of specific distances reflects similarity in amino acid sequence. High positive correlations are thus generated by related sequences. Given precautions in assessing the significance of such correlations (stringent levels of significance such as a probability of 0.001, plus transformation of r to z), subgroups of related sequences can be established. The significantly related sequences revealed by this approach were compared with those from clustering.

Finally, profiles for the sequence subgroups located among this set of 18 sequences were constructed to extract common features within and between the subgroups. Profiles can potentially be used (1) to assign functions to new sequences lacking this information, (2) search sequence databases for proteins with peroxidasic functions or proteins with components from peroxidases, and (3) provide guidelines for engineering peroxidase modification and design.

Methods

Enzyme sequence sources

The GENBANK, EMBL and Swiss Protein databases, plus relevant journals, supplied amino acid sequences plus authors/references and information on locations of active sites, oligosaccharide chain attachment points and signal peptides. Plant species are most frequently represented amongst source species, and all those available were included here. In the case of horseradish isozymes, the first documented sequence (Welinder 1979) was used. For tomato and wheat, the first isozyme listed in each original reference was used. The 8 glutathionine sequences found in the literature are virtually identical; a single sequence was included here. For *Arabidopsis* and peanut, both their two isozymes displayed some differences and both were included.

All 18 sequence references are listed in ascending length order in Table 1. The actual sequences are available in various databases. Any documented signal peptide sections were removed before alignment. The abbreviations used in Fig. 1 (dendrogram) for each of the 18 sequences are shown in the headings for the sequences.

Alignment method

Optimum alignments were calculated using a modified algorithm of Sellers (1974), and were programmed, using integer arithmetic, in compiled BASIC running on a Macintosh II. The two gap penalties W_1 and W_2 were set on the basis of preliminary tirals with a range of values. Constant penalties of 8 for W_1 and for W_2 were then used in all pairwise alignments among the 18 sequences that generated $(n^2-n)/2=153$ pairwise comparisons. Total distance between two sequences (zero for identical sequences) is either (a) the weighted number of mismatches obtained divided by the total length of the longer sequence, the measure used exclusively in analyses here or (b) the weighted number of mismatches over the final, gapped length of the longer sequence after gap insertion. The correspondence of 'distance' to 'similarity' has been demonstrated by Smith et al. (1981). The programme inputs sequences sorted for length (shortest first, longest last), and stores alignments, matching amino acids, and total distances (a and b) to disc files.

The statistical significance of total distances >0 was not tested.

Clustering techniques

The sequences were clustered using the SYSTAT statistical software package, versions 3.2 and 5.1 for the Macintosh (Wilkin-

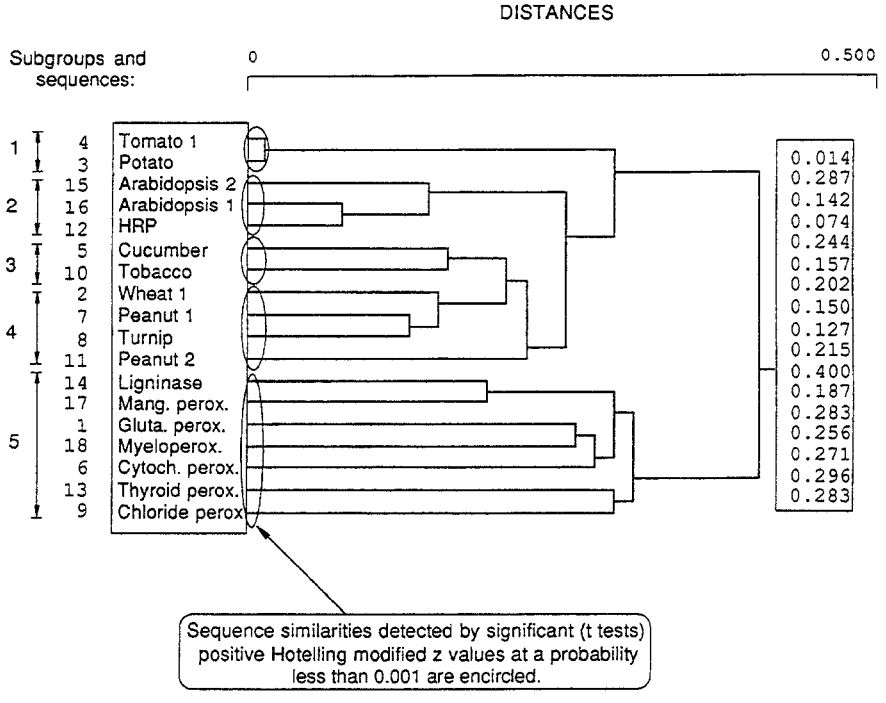

Fig. 1. Dendrogram showing relationships among 18 animal, microbial and plant peroxidase amino acid sequences. Sequence numbers as in Table 1. Branch distances shown in *box* at *right* of diagram. *Circled sequences* are related on basis of analysis of specific distances between sequences

Table 1. References for the 18 peroxidase sequences, listed in order of increasing length. Sequence numbers, 1–18, are used throughout all following tables. Wheat 1 and tomato 1 are isozymes within species. Name abbreviations shown in heading, e.g. Sequence #1, i.e. Glutathione peroxidase, is Sequence #1 ('Gluta. perox.')

Glutathione peroxidase	**Sequence #1 ('Gluta. perox.')** (*Bos primigenius taurus*, Fang et al. 1986)	205 amino acids
Wheat peroxidase 1	**Sequence #2 ('Wheat 1')** (*Triticum aestivum*, Hertig et al. 1991)	289 amino acids
Potato peroxidase	**Sequence #3 ('Potato')** (*Solanum tuberosum*, Roberts et al. 1988)	290 amino acids
Tomato peroxidase 1	**Sequence #4 ('Tomato 1')** (*Lycopersicon esculentum*, Roberts and Kolattukudy 1989)	290 amino acids
Cucumber peroxidase	**Sequence #5 ('Cucumber')** (*Cucumis sativus*, Morgens et al. 1990)	293 amino acids
Cytochrome c peroxidase	**Sequence #6 ('Cytoch. perox.')** (*Sacch. cerevisiae*, Kaput et al. 1982)	294 amino acids
Peanut peroxidase 1	**Sequence #7 ('Peanut 1')** (*Arachis hypogea*, Buffard et al. 1990)	294 amino acids
Turnip peroxidase	**Sequence #8 ('Turnip')** (*Brassica napa*, Mazza and Welinder 1980)	296 amino acids
Chloride peroxidase	**Sequence #9 ('Chloride perox.')** (*Caldariomyces fumago*, Fang et al. 1986)	300 amino acids
Tobacco peroxidase	**Sequence #10 ('Tobacco')** (*Nicotiana tabacum*, Lagrimini et al. 1987)	302 amino acids
Peanut peroxidase 2	**Sequence #11 ('Peanut 2')** (*Arachis hypogea*, Buffard et al. 1990)	306 amino acids
Horseradish peroxidase	**Sequence #12 ('HRP')** (*Armoracia rusticana*, Welinder 1979)	308 amino acids
Thyroid peroxidase	**Sequence #13 ('Thyroid perox.')** (*Sus scrofa*, Magnusson et al. 1986)	332 amino acids
Ligninase	**Sequence #14 ('Ligninase')** (*Phan. chrysosporium*, Holzbaur and Tien 1988)	344 amino acids
Arabidopsis peroxidase 2	**Sequence #15 ('Arabidopsis 2')** (*Arabidopsis thaliana*, Intapruk et al. 1991)	349 amino acids
Arabidopsis peroxidase 1	**Sequence #16 ('Arabidopsis 2')** (*Arabidopsis thaliana*, Intapruk et al. 1991)	354 amino acids
Manganese peroxidase	**Sequence #17 ('Mang. perox.')** (*Phan. chrysosporium*, Godfrey et al. 1990)	357 amino acids
Myeloperoxidase	**Sequence #18 ('Myelo. perox.')** (*Homo sapiens*, Yamada et al. 1987)	467 amino acids

son 1989). The 18 by 18 symmetrical total distance matrix was the input data; this was converted to Euclidean distances and clustered using complete linkage (farthest neighbour) procedures.

Calculation of specific distances

The calculation of the specific distances is made from the 153 total distances (computed using $W_1 = W_2 = 8$). The total distances are shown (numerals in italics) in the upper right triangle of Table 2. The procedure for specific distance calculation follows Griffing's (1956) method for computing specific combining ability values for diallel crosses lacking representatives of parents and reciprocals in the progeny generation (method 4). The application of the procudure to total distance data has been detailed elsewhere (Tyson 1991).

Specific distances were calculated for each of the 18 sequences from the total distances shown in the upper right triangle of Table 2; the resultant specific distances are shown in the lower left triangle of this table.

Correlations between specific distances of individual sequences

For two identical sequences, with a zero total distance between them, the specific combining ability/specific distance formula above produces, from a square symmetrical matrix of total distances supplied by the alignment programme, corresponding specific distances that are identical. The two identical sequences 'interact' in exactly the same way with all other sequences; sequences differing to any extent from one another (total distance >0) do not, and their patterns of specific distances will not be completely correlated. A positive linear similarity can be discerned through the sign/size of the simple (Pearson) correlation coefficient r. Subgroups of sequences can thus be established. Two such related sequences will generally display a correspondingly high proportion of amino acid matches, but may not nec-

Table 2. Distances (upper right, italics); specific distances (lower left). Sequence numbers as in Table 1

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0	*−0.753*	*−0.304*	*−0.288*	*−0.773*	*0.985*	*−0.808*	*−0.793*	*0.981*
2	0.059	0	*0.070*	*0.046*	*0.636*	*−0.785*	*0.950*	*0.925*	*−0.776*
3	0.058	0.032	0	*0.978*	*0.051*	*−0.294*	*0.198*	*0.224*	*−0.299*
4	0.063	0.026	−0.563	0	*0.013*	*−0.267*	*0.202*	*0.243*	*−0.316*
5	0.078	−0.014	0.012	0.003	0	*−0.756*	*0.624*	*0.549*	*−0.744*
6	−0.108	0.063	0.046	0.055	0.071	0	*−0.840*	*−0.829*	*0.993*
7	0.086	−0.181	−0.026	−0.026	−0.018	0.087	0	*0.977*	*−0.845*
8	0.068	−0.134	−0.035	−0.031	−0.023	0.076	−0.186	0	*−0.840*
9	−0.133	0.076	0.074	0.068	0.060	−0.099	0.104	0.086	0
10	0.067	−0.048	0.030	0.023	−0.141	0.052	−0.049	−0.037	0.066
11	0.009	−0.041	0.033	0.031	−0.018	0.027	−0.054	−0.072	0.040
12	0.085	−0.049	0.067	0.070	−0.073	0.093	−0.036	−0.027	0.101
13	−0.147	0.065	0.047	0.039	0.038	−0.123	0.084	0.067	−0.140
14	−0.105	0.065	0.057	0.060	0.033	−0.099	0.087	0.079	−0.133
15	0.053	−0.022	0.016	0.027	−0.052	0.031	−0.033	0.011	0.045
16	0.061	−0.017	0.051	0.057	−0.042	0.047	−0.008	0.006	0.039
17	−0.099	0.064	0.051	0.044	0.042	−0.110	0.091	0.082	−0.119
18	−0.097	0.056	0.051	0.055	0.043	−0.110	0.078	0.071	−0.134

	10	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
1	*−0.819*	*−0.624*	*−0.621*	*0.971*	*0.830*	*−0.699*	*−0.580*	*0.841*	*0.987*
2	*0.706*	*0.862*	*0.474*	*−0.758*	*−0.656*	*0.514*	*0.443*	*−0.672*	*−0.768*
3	*0.008*	*−0.069*	*−0.097*	*−0.227*	*−0.226*	*−0.063*	*−0.113*	*−0.199*	*−0.305*
4	*−0.014*	*−0.084*	*−0.103*	*−0.247*	*−0.208*	*−0.040*	*−0.111*	*−0.211*	*−0.288*
5	*0.913*	*0.622*	*0.656*	*−0.830*	*−0.695*	*0.682*	*0.654*	*−0.689*	*−0.776*
6	*−0.839*	*−0.600*	*−0.582*	*0.959*	*0.843*	*−0.719*	*−0.600*	*0.828*	*0.971*
7	*0.689*	*0.787*	*0.445*	*−0.807*	*−0.712*	*0.451*	*0.399*	*−0.704*	*−0.831*
8	*0.677*	*0.731*	*0.356*	*−0.778*	*−0.672*	*0.476*	*0.333*	*−0.666*	*−0.797*
9	*−0.796*	*−0.582*	*−0.536*	*0.957*	*0.824*	*−0.679*	*−0.590*	*0.842*	*0.972*
10	0	*0.653*	*0.685*	*−0.799*	*−0.726*	*0.734*	*0.631*	*−0.717*	*−0.788*
11	−0.058	0	*0.436*	*−0.574*	*−0.479*	*0.295*	*0.297*	*−0.543*	*−0.615*
12	−0.063	0.013	0	*−0.683*	*−0.662*	*0.934*	*0.929*	*−0.682*	*−0.642*
13	0.066	0.033	0.097	0	*0.830*	*−0.710*	*−0.599*	*0.825*	*0.985*
14	0.063	0.028	0.095	−0.080	0	*−0.743*	*−0.610*	*0.995*	*0.829*
15	−0.048	−0.021	−0.175	0.076	0.123	0	*0.937*	*−0.768*	*−0.664*
16	−0.048	0.011	−0.362	0.089	0.125	−0.205	0	*−0.611*	*−0.542*
17	0.075	0.018	0.096	−0.083	−0.327	0.116	0.133	0	*0.821*
18	0.052	0.021	0.067	−0.127	−0.072	0.056	0.065	−0.075	0

cessarily do this in all cases. The correlation of one sequence's specific distances with those of another sequence was calculated after removing, from each sequence's column of specific distances, (1) the zero specific distance of the sequence from itself and (2) the specific distance between the two sequences being correlated. This eliminated bias in the r values; the remaining $(18-2)=16$ specific distance pairs generating the r value had 14 degrees of freedom (df). The r values are shown (numerals in italics) in the upper right triangle of Table 3. The application of correlation to the specific distance patterns has also been detailed elsewhere (Tyson 1991).

The r values in the upper right triangle of Table 3 were transformed to z to deal with their distribution problems, and since the sample sizes were relatively small, the additional Hotelling (1953) modification of z (Sokal and Rohlf 1981) was employed, plus standard errors calculated as $\sqrt{(1/(n-1))}$ where $n=16$. The lower left triangle of Table 3 shows the Hotelling z values. A tabulated t with 14 df at probability 0.001 ($t=4.140$) was used. The siginficant t's for positive z values are shown in bold type in the square symmetrical 18 by 18 matrix of Table 4, which allows sequence relationshps to be read from each column or row.

Profiles of related sequences

For two or more closely related sequences of similar length, a conventional first approach to discerning conserved amino acids and consistent motifs is to superimpose them. A minimal number of gaps is then inserted by eye to maximise matching across all sequences. A more controlled alternative uses the optimum alignment programme to provide a computed pairwise alignment and a list of amino acid matches including inserted gaps. Matches plus gaps for one pairwise alignment can be compared with that for any other calculated at the same W_1 and W_2 values. Common sequence features can be assessed from (1) the aligned sequences or (2) the list of matching amino acids.

In (1) the pairs of aligned sequences (including inserted gaps) are exactly superimposed. Variation among elements (amino acids and/or gaps) can then be examined at each position and invariant amino acids and the amino acid frequency distribution at each position from the NH_2 to the COOH terminus obtained.

In (2) the matching amino acids detected by alignments for each sequence pair are listed as an amino acid series without any gaps. The lists of matching amino acids generated by all pairwise alignments are superimposed. Gaps are then reinserted visually

Table 3. *r* values (upper right, italics); Hotelling *z* values (lower left). Sequence numbers as in Table 1

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0	*−0.753*	*−0.304*	*−0.288*	*−0.773*	*0.985*	*−0.808*	*−0.793*	*0.981*
2	−0.918	0	*0.070*	*0.046*	*0.636*	*−0.785*	*0.950*	*0.925*	*−0.776*
3	−0.293	0.065	0	*0.978*	*0.051*	*−0.294*	*0.198*	*0.224*	*−0.299*
4	−0.277	0.043	2.125	0	*0.013*	*−0.267*	*0.202*	*0.243*	*−0.316*
5	−0.964	0.703	0.048	0.012	0	*−0.756*	*0.624*	*0.549*	*−0.744*
6	2.319	−0.992	−0.283	−0.255	−0.924	0	*−0.840*	*−0.829*	*0.993*
7	−1.052	1.726	0.188	0.191	0.684	−1.146	0	*0.977*	*−0.845*
8	−1.013	1.524	0.213	0.232	0.578	−1.112	2.090	0	*−0.840*
9	2.196	−0.970	−0.288	−0.305	−0.898	2.688	−1.164	−1.147	0
10	−1.082	0.823	0.007	−0.013	1.455	−1.142	0.792	0.771	−1.019
11	−0.684	1.221	−0.065	−0.078	0.682	−0.649	0.998	0.871	−0.623
12	−0.680	0.481	−0.091	−0.097	0.735	−0.623	0.447	0.348	−0.560
13	1.994	−0.928	−0.216	−0.236	−1.115	1.819	−1.049	−0.976	1.798
14	1.114	−0.736	−0.215	−0.197	−0.804	1.155	−0.834	−0.763	1.097
15	−0.810	0.531	−0.059	−0.037	0.781	−0.848	0.454	0.484	−0.775
16	−0.620	0.444	−0.106	−0.104	0.733	−0.648	0.395	0.323	−0.633
17	1.151	−0.762	−0.188	−0.200	−0.792	1.110	−0.820	−0.753	1.153
18	2.387	−0.951	−0.294	−0.276	−0.970	1.983	−1.117	−1.023	2.002

	10	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
1	*−0.819*	*−0.624*	*−0.621*	*0.971*	*0.830*	*−0.699*	*−0.580*	*0.841*	*0.987*
2	*0.706*	*0.862*	*0.474*	*−0.758*	*−0.656*	*0.514*	*0.443*	*−0.672*	*−0.768*
3	*0.008*	*−0.069*	*−0.097*	*−0.227*	*−0.226*	*−0.063*	*−0.113*	*−0.199*	*−0.305*
4	*−0.014*	*−0.084*	*−0.103*	*−0.247*	*−0.208*	*−0.040*	*−0.111*	*−0.211*	*−0.288*
5	*0.913*	*0.622*	*0.656*	*−0.830*	*−0.695*	*0.682*	*0.654*	*−0.689*	*−0.776*
6	*−0.839*	*−0.600*	*−0.582*	*0.959*	*0.843*	*−0.719*	*−0.600*	*0.828*	*0.971*
7	*0.689*	*0.787*	*0.445*	*−0.807*	*−0.712*	*0.451*	*0.399*	*−0.704*	*−0.831*
8	*0.677*	*0.731*	*0.356*	*−0.778*	*−0.672*	*0.476*	*0.333*	*−0.666*	*−0.797*
9	*−0.796*	*−0.582*	*−0.536*	*0.957*	*0.824*	*−0.679*	*−0.590*	*0.842*	*0.972*
10	0	*0.653*	*0.685*	*−0.799*	*−0.726*	*0.734*	*0.631*	*−0.717*	*−0.788*
11	0.731	0	*0.436*	*−0.574*	*−0.479*	*0.295*	*0.297*	*−0.543*	*−0.615*
12	0.784	0.437	0	*−0.683*	*−0.662*	*0.934*	*0.929*	*−0.682*	*−0.642*
13	−1.028	−0.612	−0.782	0	*0.830*	*−0.710*	*−0.599*	*0.825*	*0.985*
14	−0.863	−0.488	−0.746	1.115	0	*−0.743*	*−0.610*	*0.995*	*0.829*
15	0.878	0.284	1.591	−0.830	−0.897	0	*0.937*	*−0.768*	*−0.664*
16	0.696	0.286	1.556	−0.647	−0.664	1.611	0	*−0.611*	*−0.542*
17	−0.844	−0.569	−0.779	1.101	2.823	−0.952	−0.664	0	*0.821*
18	−0.999	−0.670	−0.712	2.317	1.113	−0.750	−0.568	1.089	0

so that consistent features at the same relative position in the superimposed lists can form the sequences' profile. This method is appropriate where sequences are variable in length. Its use implies that the order of matching elements is more important than their absolute position. This procedure was used here.

Results

Sequence comparisons

The 18 sequences being compared are shown in Table 1. The 153 alignments at $W_1 = 8$, $W_2 = 8$ of these sequences generated the total distances shown in the upper right triangle (italics) of Table 2, with the specific distances calculated from these distances appearing in the lower left triangle.

The total distance data in the upper right triangle of Table 2 were used to compute the dendrogram for the 18 sequences in Fig. 1; the sequence abbreviations used are shown in the Table 1 heading for each sequence, e.g. sequence #1 ('Gluta. perox.'). The Fig. 1 dendrogram displays a clear demarcation between plant peroxidases in the upper part and animal and microbial peroxidases in the lower. The branching distances boxed on the right of Fig. 1, as well as the positions of branches on the branch distance scale at the top, suggest that the relationships among plant peroxidases are generally tighter than those among the seven animal/microbial sequences.

Amongst the latter seven sequences, ligninase (sequence 14) and manganese peroxidase (sequence 17), both sequenced in *Phanerochaete chrysosporium*, appear to be the most closely related, with lesser similarities between sequences 1, 18 and 6, and between 19 and 13. This is confirmed by inspection of the specific distance analysis results (see below). Notwithstanding the length difference between sequences 1 and 18, they are related. Individual alignments are examined below.

Table 4. *t* tests of Hotelling z values. Significant positive z's in bold figures. *t* for 14 df at probability 0.001 = 4.14. Sequence numbers as in Table 1

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0	−3.554	−1.133	−1.072	−3.735	**8.982**	−4.075	−3.924	**8.504**
2	−3.554	0	0.253	0.116	2.723	−3.842	**6.684**	**5.901**	−3.757
3	−1.133	0.253	0	**8.231**	0.185	−1.094	0.727	0.824	−1.114
4	−1.072	0.166	**8.231**	0	0.047	−0.989	0.740	0.898	−1.182
5	−3.735	2.723	0.185	0.047	0	−3.578	2.651	2.237	−3.480
6	**8.982**	−3.842	−1.094	−0.989	−3.578	0	−4.439	−4.305	**10.411**
7	−4.075	**6.684**	0.727	0.740	2.651	−4.439	0	**8.094**	−4.508
8	−3.924	**5.901**	0.824	0.898	2.237	−4.305	**8.094**	0	−4.442
9	**8.504**	−3.757	−1.114	−1.182	−3.480	**10.411**	−4.508	−4.442	0
10	−4.189	3.187	0.028	−0.050	**5.637**	−4.423	3.068	2.987	−3.946
11	−2.648	**4.730**	−0.250	−0.304	2.640	−2.514	3.866	3.374	−2.412
12	−2.635	1.865	−0.353	−0.374	2.847	−2.412	1.731	1.347	−2.168
13	**7.723**	−3.595	−0.835	−0.913	−4.317	**7.044**	−4.062	−3.778	**6.965**
14	**4.315**	−2.850	−0.833	−0.765	−3.112	**4.474**	−3.230	−2.956	**4.250**
15	−3.138	2.058	−0.230	−0.143	3.023	−3.285	1.759	1.876	−3.000
16	−2.402	1.721	−0.410	−0.402	2.839	−2.511	1.530	1.253	−2.452
17	**4.456**	−2.950	0.728	−0.776	−3.068	**4.298**	−3.175	−2.917	**4.467**
18	**9.244**	−3.683	−1.140	−1.071	−3.758	**7.678**	−4.324	−3.964	**7.755**

	10	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
1	−4.189	−2.648	−2.635	**7.723**	**4.315**	−3.138	−2.402	**4.456**	**9.244**
2	3.187	**4.730**	1.865	−3.595	−2.850	2.058	1.721	−2.950	−3.683
3	0.028	−0.250	−0.353	−0.835	−0.833	−0.230	−0.410	−0.728	−1.140
4	−0.050	−0.304	−0.374	−0.913	−0.765	−0.143	−0.402	−0.776	−1.071
5	**5.637**	2.640	2.847	−4.317	−3.112	3.023	2.839	−3.068	−3.758
6	−4.423	−2.514	−2.412	**7.044**	**4.474**	−3.285	−2.511	4.298	7.678
7	3.068	3.866	1.731	−4.062	−3.230	1.759	1.530	−3.175	−4.324
8	2.987	3.374	1.347	−3.778	−2.956	1.876	1.253	−2.917	−3.964
9	−3.946	−2.412	−2.168	**6.965**	**4.250**	−3.000	−2.452	**4.467**	**7.755**
10	0	2.830	3.038	−3.983	−3.342	3.400	2.695	−3.268	−3.869
11	2.830	0	1.691	−2.370	−1.889	1.100	1.106	−2.203	−2.595
12	3.038	1.691	0	−3.030	−2.888	**6.162**	**6.027**	−3.018	−2.758
13	−3.983	−2.370	−3.030	0	**4.320**	−3.216	−2.504	**4.263**	**8.973**
14	−3.342	−1.889	−2.888	**4.320**	0	−3.473	−2.571	**10.935**	**4.310**
15	3.400	1.100	**6.162**	−3.216	−3.473	0	**6.240**	−3.688	−2.903
16	2.695	1.106	**6.027**	−2.504	−2.571	**6.240**	0	−2.572	−2.201
17	−3.268	−2.203	−3.018	**4.263**	**10.935**	−3.688	−2.572	0	**4.218**
18	−3.869	−2.595	−2.758	**8.973**	**4.310**	−2.903	−2.201	**4.218**	0

Among the plant peroxidases in the upper part of Fig. 1, at least four subgroups are suggested by the dendrogram branches. The marked length difference between the 2 peanut isozymes is reflected in the isolation of peanut 2 from sequences 2, 5, 7 (peanut 1), 8 and 10. The 2 *Arabidopsis* isozymes are much closer to one another. The extremely close relationship between tomato 1 and potato (Solanaceae), and between cucumber and tobacco, were detected in preliminary studies of the plant peroxidases that included all available isozyme sequences. The four subgroups anticipated among the plant sequences from Fig. 1 are complemented by possibly three subgroups among the seven animal/microbial sequences.

Specific distance analysis

The specific distances calculated from the distance data are shown in the lower left triangle of Table 2. Pearson correlation coefficients (*r*) calculated from all pairwise comparisons among the columns of the specific distance matrix are shown in the upper right triangle (italics) of Table 3. The corresponding z values, using the Hotelling (1973) modification for small sample sizes, are shown in the lower left triangle of Table 3. The *t* tests of the Hotelling z values are shown in Table 4, and for positive z values those exceeding the theoretical *t* value for a 0.001 probability with 14 degrees of freedom (4.140) are shown in bold type. The Fig. 1 dendrogram has been completed by the incorporation of these Table 4 specific distance results; sequences significantly related are circled.

Among the animal and microbial sequences the significant *t*'s indicate that these seven sequences form a single subgroup (sequences 1, 6, 9, 13, 14, 17 and 18), in contrast to the plant sequences where four subgroups occur. These are sequences 3 and 4 (subgroup 1), 12, 15 and 16 (subgroup 2), 5 and 10 (subgroup 3), and 2, 7, 8

and 11 (subgroup 4). Subgroup 5 contains sequences 1, 6, 9, 13, 14, 17 and 18.

All *t* tests of z values in Table 4 use the same standard error; each *t* value therefore reflects the size of the corresponding z and hence *r*. The average of the significant positive *t* test values for the 11 plant peroxidases (6.46) is essentially the same as that for the 7 others (6.55). However, the averages of all *t*'s in Table 4 for sequences 1, 6, 9, 13, 14, 17 and 18 is 0.44, whereas the average for the 11 plant peroxidases is 0.16 due to the low or negative values in their columns. The animal/microbial sequences are more homogeneous in terms of their behaviour when aligned with all other sequences in the set of 18.

There is general agreement between dendrogram and specific distance results as shown by circled sequences in Fig. 1. Table 4 shows that sequence 12 is related to the same degree to sequences 15 and 16 (the two *Arabidopsis* isozymes), whereas Fig. 1 suggests that 12 and 16 are more closely related to each other than to 15. Although the 2 peanut isozymes (sequences 7 and 11) are not significantly related, the z value (+3.866) is large, positive and approaches 4.140. Sequence 2 is, however, related to sequences 7, 8 and 11; thus these four sequences are linked in a chain to create subgroup 4.

For the plant peroxidases in subgroups 1–4, Fig. 1 branch distance data was used, and the average of their ten branches was found to be 0.1612 compared to the 0.2627 average for the animal/microbial peroxidases in subgroup 5. The branch distances within the circled subgroups 1, 2, 3 and 4 are, however, relatively low compared to distances for branches connecting these subgroups. The converse situation exists in subgroup 5, so that no contradiction exists between the clustering and specific distance analyses.

Individual alignments

Sequences 1 and 18 are related (Table 4); most matches in their alignment are collected into short motifs with large intervening stretches, suggesting either that sequence 1 may be formed by deletions of parts of 18 or that 18 contains parts of 1 with numerous additional amino acids inserted.

The matches in the alignment of sequences 6 and 18 are scattered more or less uniformly throughout, with no evidence of motifs. For sequence 1 aligned against 6 there is again a suggestion that matches are collected and motifs occur. Table 5 shows the alignment of ligninase (sequence 14) and manganese peroxidase (sequence 17). Here there is a much higher proportion of matches and clear evidence of conserved areas for 2 sequences. The proportions of matches and inserted gaps in this particular alignment in group 5 is approximately of the same order as those seen for the alignments within each of the groups 1–4, except for the extremely close resemblance of tomato and potato.

Examination of individual alignments leads into the question of summarising the sequence relationships and extracting the essential features of this set of sequences.

Profiles

The lists of matching amino acids generated by optimum alignment were used to construct profiles for individual subgroups.

Subgroups 1, 2, 3 and 4. Profiles were firstly constructed for each of subgroups 1–4. For subgroups 1 and 3 this was the list of matching amino acids. For the three sequences in subgroup 2, there were three such lists of matches corresponding to the three pairwise sequence combinations. Gaps were removed from these three lists, which were then superimposed. Conserved areas of identical amino acids were then readily discernable, although vertically out of line in some cases. They were lined up by visually reinserting gaps. For groups of conserved amino acids this was straightforward; for single amino acids clues to correct ordering were provided by their relations to these groups. The result emphasised the consistent ordering of conserved elements among sequences within subgroups. Where an amino acid was completely consistent over the three realigned lists for subgroup 2, this element was placed in the subgroup's profile. This was a conservative approach to extract the most essential elements by providing a stringent screening for essential amino acids from the NH_2 to the COOH terminus. The profile for subgroup 3 was constructed in a similar fashion; the 4 sequences (2, 7, 8 and 11) provided six amino acid match lists. Exclusion of the 'outlying' sequence 11 had virtually no effect on the end result.

Secondly, the four resultant profiles for subgroups 1–4 were superimposed and are shown in Fig. 2. In this diagram the profiles with approximately, 300 elements each have been folded into consecutive blocks in the directions indicated by arrows and numbering. The order of amino acids in the profiles corresponds to the order in the original sequences, but the numbers 1–320 are merely inserted for convenience in referring to a particular section within the profiles. The first four rows of Fig. 2 depict the four superimposed and realigned profiles. For elements 1–50 the first four rows correspond, respectively, to profiles from sequences 3 and 4, sequences 12, 15 and 16, sequences 5 and 10 and sequences 2, 7, 8 and 11. The same ordering holds for elements 51–100, and so on. The final (fifth) row of 1–50 contains the consensus for the four profiles above. The same organisation holds for 51 onwards.

Where three of the four profiles display consistent features, e.g. 'FY' and 'CPV' at the NH_2 terminus (7–12), the respective amino acids are highlighted and incorporated in the final profile. The fifth row ('Consensus')

Table 5. Alignment of sequence 14 (ligninase) against sequence 17 (manganese peroxidase). Asterisks indicate exact matches, dashes indicate gaps inserted to maximise matches using $W_1 = W_2 = 8$. S1 is sequence 14; S2 is sequence 17

	5	10	15	20	25	30	35	40	45	50
S1	ATCSN	GKTVG	DASCC	AWFDV	LDDIQ	QNLFH	GGQCG	AEAHE	SIRLV	FHDSI
	* *	* *	* * *	* *	*	*	* *	* * *	* * *	* * * *
S2	AVCPD	GTRVS	HAACC	A-FIP	LAQDL	QETIF	QNECG	EDAHE	VIRLT	FHDAI
	55	60	65	70	75	80	85	90	95	100
S1	AISPA	MEAQG	KFGGG	GADGS	IMIFD	DIETA	FHPNI	GLDEI	VKLQK	PFVQK
	* * *	* *	* *	* * * * *	*	*	* *	* *	*	* * * *
S2	AIS--	-RSQG	PKAGG	GADGS	MLLFP	TVEPN	FSANN	GIDDS	VNNLI	PFMQK
	105	110	115	120	125	130	135	140	145	150
S1	H-GVT	PGDFI	AFAGR	VALSN	CPGAP	QMNFF	TGRAP	ATQPA	PDGLV	PEPFH
	*	*	* * *	* * * * *	* * * * *	*	* *	* *	* * *	* * *
S2	HNTIS	AADLV	QFAGA	VALSN	CPGAP	RLEFL	AGRPN	KTIAA	VDGLI	PEPQD
	155	160	165	170	175	180	185	190	195	200
S1	TVDQI	INRVN	DAGEF	DELEL	VWMLS	AHSVA	AVNDV	DPTVQ	GLPFD	STPGI
	* *	*	* * * *	*	* *	* * * *	*	* *	* * *	* * *
S2	SVTKI	LQRFE	DAGGF	TPFEV	VSLLA	SHSVA	RADKV	DQTID	AAPFD	STPFT
	205	210	215	220	225	230	235	240	245	250
S1	FDSQF	FVETQ	LRGTA	FPGSG	GNQGE	VESPL	P----	---GE	IRIQS	DHTIA
	* * *	*	* *	* * * *	* * *	* * * *	*	* *	* * *	* *
S2	FDTQV	FLGVL	LKGVG	FPGSA	NNTGE	VASPL	PLGSG	SDTGE	MRLQS	DFALA
	255	260	265	270	275	280	285	290	295	300
S1	RDSRT	ACEWQ	SFVNN	QSKLV	DDFQF	IFLAL	TQLGQ	DPNAM	TDCSD	VIPQS
	* * *	* * *	* * *	*	*	*	* *	*	* * * *	* *
S2	HDPRT	ACIYQ	GFVNE	QAFMA	ASFRA	AMSKL	AVLGH	NRNSL	IDCSD	VVPVP
	305	310	315	320	325	330	335	340	345	350
S1	KPIPG	NLPFS	FFPAG	KTIKD	VEQAC	AETPF	PTLTT	LPGPE	TSVQR	IPP--
	* * *	*	* * *	*	* *	*	* * * * *	* *	*	*
S2	KPATG	-QP-A	MFPAS	TGPQD	LELSC	PSERF	PTLTT	QPGAS	QSLIA	HCPDG
	355	360								
S1	---PP	G----	--A							
	*	*	*							
S2	SMSCP	GVQFN	GPA							

Number of gaps in sequences 14 and 17 are 19 and 6, respectively. Number of matches = 164. Total distance = 0.5613

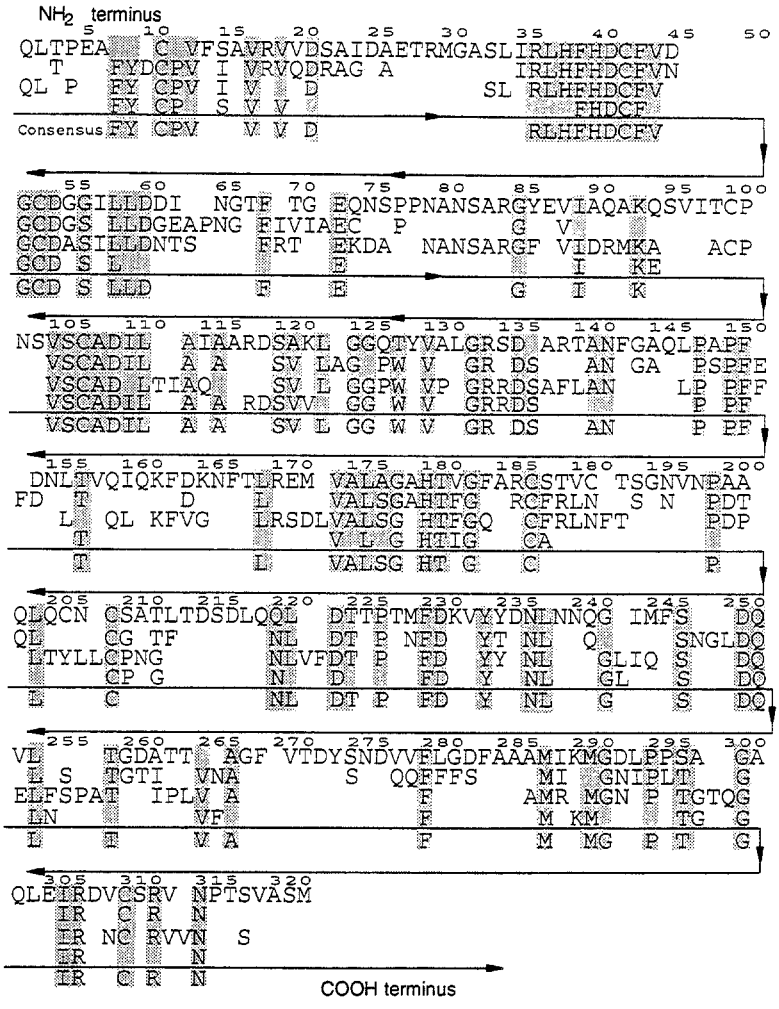

Fig. 2. Profiles of subgroups 1, 2, 3 and 4, and profile for all 11 plant peroxidases (consensus). All profiles are folded as indicated by *arrows* and *numbering* to fit into seven consecutive blocks within the page. Amino acids consistently found in three of the four subgroup profiles are *highlighted*. Total for 75% consistency = 98 amino acids; total for 100% consistency = 54 amino acids. *Row 1* contains the profile for subgroup 1 (sequences 3, 4), *row 2* profile for subgroup 2 (sequences 12, 15, 16), *row 3* profile for subgroup 3 (sequences 5, 10), *row 4* profile for subgroup 4 (sequences 2, 7, 8, 11). *Row 5* contains consensus profile for all subgroups

displays the overall plant peroxidase profile. Given an average length of 300 amino acids, the overall plant peroxidase profile contains 98 elements, or 33% of the average length. Of the 98, 54 are completely consistent across all four profiles. The 98 amino acids contain six cysteines (positions 10, 41, 52, 105, 185 and 207) involved with three disulphide bridges, which are completely consistent across all profiles, plus cysteines at 98 and 307, which may be involved with a fourth bridge. There are also two completely consistent histidines (proximal, 39, and distal, 178), which locate the haem group.

Three sections each contain more than three consecutive and consistent amino acids, namely 'RLHFHDCFV' (35–43), 'VSCADIL' (102–109) and 'VALSG' (172–176), which is extended with '–HT–G' (177–191) and holds the second H ligand. The 35–43 section is a highly conserved box, as is the 172–188 section. There are four asparagine residues (positions 140, 218 235 and 313) These asparagines all occur in limited conserved areas, unlike the haem ligands, and are potential attachment sites for oligosaccharide chains. These basis features are now well established in reviews of plant peroxidase amino acid sequences. What appears from these profiles, however, is the high degree of variability occurring between the evenly scattered islands of consistent elements that conserve the essential three-dimensional structure of the enzyme.

The scoring of exact matches only in alignments and the requirement of complete consistency for inclusion in a subgroup profile retains the most important amino acids. The final plant peroxidase profile (75% consistent), which appears in the fifth row of Fig. 2, is not neccessarily the unique profile maximising conserved elements, but certainly comes close to the optimum. It can then be compared with profiles for peroxidases of other types and/or from other sources and confidently examined for clues to structural characteristics. With respect to sequence comparisons in general, Degrado et al. (1989), and Lesk and Chothia (1982) have suggested that proteins with fewer than 10% of the same residues at comparable positions in the protein chain can have remarkably similar structures.

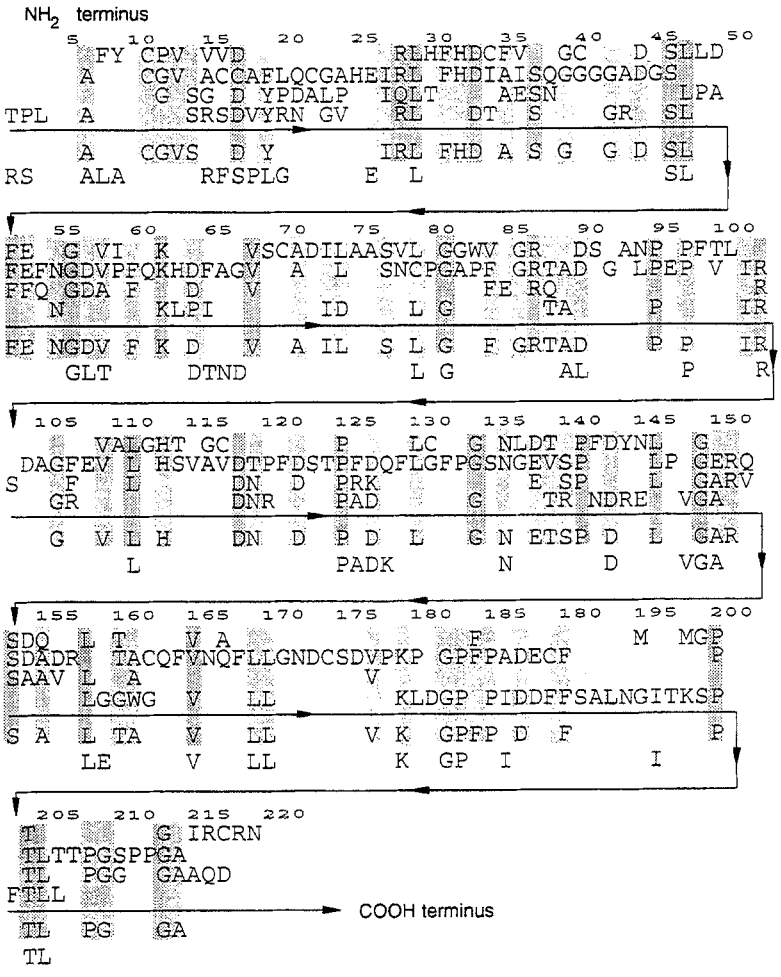

Fig. 3. Profiles of sequences in subgroup 5 plus profile for all plant peroxidases and profile for all peroxidases in subgroups 1–5 (except sequence 1). All profiles are folded as indicated by *arrows* and *numbering* to fit into five consecutive blocks within the page. Amino acids consistently found in two or three of the first four rows of profiles are *highlighted* and shown in row 5. Total for 50% consistency = 90 amino acids; total for 75% consistency = 29 amino acids. *Row 1* contains consensus profile from Fig. 2 (row 5) profile for sequences 14 and 17, *row 3* profile for sequences 9 and 13, *row 4* profile for sequences 6 and 18. *Row 5* contains final profile for rows 1–4. *Row* 6 contains profile for sequences 1 and 6, 1 and 18, and 6 and 18, i.e. all alignments between 1, 6 and 18

Subgroup 5 and the profile for all subgroups. The sequences in subgroup 5 were handled in two stages as before. To reduce the 21 sequence combinations required for the 7 sequences, these were treated as three separate subgroups, and the lists of matching amino acids for alignments 14 by 17, 9 by 13, and 6 by 18 were superimposed. The plant peroxidase profile was also included and forms the top row of Fig. 3. Row two, contains the profile for the sequence 14 by 17 alignment; row three, the 9 by 13; row four, the 6 by 18. This diagram folds the superimposed sequences as in Fig. 2, indicated by the arrows and numbering. Glutathione peroxidase (sequence 1) was omitted from the 1 by 6 by 18 profile construction since it lacked many of the components appearing in other profiles. The final, fifth row contains the consensus and is the profile for all peroxidases in this study.

Because matches among the subgroup 5 alignments are fewer than those for subgroups 1–4, amino acids occurring in two as well as in three of the four rows are also included in the final profile. The 50% consistency is lightly highlighted; the 75% consistency is distinguished by heavier highlighting. This results in 90 elements appearing in the final profile for peroxidases. The 50% consistency of amino acids is randomly distributed over the four rows. The effect of sequence 1 on the profile for 6 and 18 is shown in row six of the diagram. This supplies the profile for matches in 1 by 6, 1 by 18 and 6 by 18, and contains only 50 elements. Large pieces are clearly missing in comparison to profiles in rows 1, 2 and 3, given that the ordering of these 50 amino acids for 1, 6 and 18 is correct.

The final profile of 88 elements in row five shows that the main features detected for the 11 plant peroxidases reappear in the animal/microbial peroxidases and the final profile. A major part of the first conserved box at 35–43 is seen in the final profile at 26–32. The second box is only partially represented at 107–111, but the conserved area at the COOH terminus (293–300 in Fig. 2) is seen at 198–213 in the final profile in Fig. 3. As for the plant peroxidases in Fig. 2, this final Fig. 3 profile emphasises the even distribution of conserved elements and motifs throughout the length of the enzyme protein. No amino acid substitutions or conservative replace-

ments were involved in the establishment of these constant features of linear peroxidase sequences; in contrast, see Fujiyama et al. (1990) for comparisons among four sequences. Notably absent from the final profile in Fig. 3 are the conserved cysteines appearing in Fig. 2. The one cysteine (Fig. 3, position 10) suggests that disulphide bridges vary in position, reflecting variations in three-dimensional structure for functional reasons. Asparagine (N linkages) sites for carbohydrate attachment are known to vary considerably from one peroxidase to another.

The profile for sequence 14 by 17 in row two in Fig. 3 contains more elements than the plant peroxidase profile in row one. This is expected since sequences 14 and 17 represent functionally similar enzyme proteins (ligninase and manganese peroxidase) generated from the same organism. Row three (profile for sequences 9 and 13, chloride and thyroid peroxidases) contains fewer elements for the opposite reasons.

Conclusions

The first main aim of this study was to determine relationships among a number of amino acid sequences from one family of enzyme proteins. An optimum alignment was used to generate a symmetrical distance matrix for currently available peroxidase amino acid sequences, thus generating a dendrogram summarising relationships among the 18 sequences by conventional clustering procedures. From the distance data, specific distances of individual sequences from each other were obtained. Comparisons of the patterns of specific distances for the 18 sequences revealed similarities and differences between these amino acid sequences. Dendrogram and specific distance analyses agreed well, the latter enabling the sequences to be allocated to subgroups and the significance of sequence relationships to be determined. These combined procedures placed recognition of relationships among a large set of sequences on a more controlled footing than possible by purely visual inspection. They avoid the assumptions required, for example, by the variously produced distance matrices employed by Dufton and Rochat (1985) in clustering to evaluate scorption toxin relationships.

The second main aim of the study was to couple the results from clustering and subgrouping with the construction of profiles that contained the essential elements for each of the five subgroups and the amalgamation of these into a single final profile summarising the essential features of all peroxidases scanned in this study. This final profile for the complete set of 18 animal, microbial and plant peroxidases agreed in many respects with sequence comparisons made by other authors. The features seen among plant peroxidases are repeated in peroxidases of more varied functions from more widely differing sources. Cysteine residues are in many cases conserved. The restriction to excact matching during alignment and complete consistency in initial profiling expands the generality of the conserved elements in the final profile. There is an unavoidable subjective facet to the visual realignments made with the lists of matching amino acids to construct profiles, but this is offset by the use of match lists and groupings produced by conservative and repeatable numerical methods.

The determination of evolutionary relationships among sequences, in contrast to their current relationships, requires numerous assumptions and has not been pursued here. Of more immediate interest will be the search for peroxidase sequence features correlated with enzyme functional adaptations and differences in natural populations. It would be valuable, for example, to unravel the sequence features that might confer tolerance to extreme or unusual environmental conditions and thus open the way to engineered improvements in this versatile and important enzyme family. The approaches used in this study show promise in allowing sequence information to be sorted to recognise such adaptational characteristics.

Acknowledgements. The work reported here was financed in part by an operating grant from the Natural Sciences and Engineering Research Council of Canada, to whom thanks are extended.

References

Buffard D, Breda C, van Huystee R, Asemota O, Piere M, Ha D, Esnault R (1990) Molecular cloning of complementary DNAs encoding two cationic peroxidases from cultivated peanut cells. Proc Natl Acad Sci USA 87:8874–8878

Degrado W, Wasserman Z, Lear J (1989) Protein design, a minimalist approach. Science 243:622–628

Dufton M, Rochat H (1984) Classification of scorpion toxins according to amino acid composition and sequence. J Mol Evol 20:120–127

Fang G, Kenigsberg P, Axley M, Nuell M, Hager L (1986) Cloning and sequencing of chloroperoxidase cDNA. Nucleic Acids Res 14:8061–8071

Fujiyama K, Takemura H, Shibayama S, Kobayashi K, Choi J-K, Shinmyo A, Takano M, Yamada Y, Okada H (1988) Structure of the horseradish peroxidase isozyme C genes. Eur J Biochem 173:681–687

Fujiyama K, Takemura H, Shinmyo A, Okada H, Takano M (1990) Genomic DNA structure of two new horseradish-peroxidase-encoding genes. Gene 89:163–169

Gaspar TH, Penel C, Thorpe T, Greppin H (1980) Peroxidases 1970–1980. A survey of their biochemical and physiological roles in higher plants. University of Geneva, Switzerland, pp 1–324

Godfrey B, Mayfield M, Brown J, Gold M (1990) Characterisation of a gene encoding a manganese peroxidase from *Phanerochaete chrysosporium*. Gene 93:119–124

Griffing B (1956) Concept of general and specific combining ability relation to diallel crossing sytems. Aust J Biol Sci 9:463–493

Gunzler W, Steffens G, Grossmann A, Kim S, Otting F, Wendel A, Flohe L (1984) The amino-acid sequence of bovine glutathione peroxidase. Hoppe-Seyler's Z Physiol Chem 365:195–212

Harvey P, Schoemaker H, Bowen R, Palmer J (1985) Single-electron transfer processes and the reaction mechanism of enzymic degradation of lignin. FEBS Lett 183:13–16

Hertig C, Rebmann G, Bull J, Mauch F, Dudler R (1991) Sequence and tissue-specific expression of a putative peroxidase gene from wheat (*Triticum aestivum* L.). Plant Mol Biol 16: 171–174

Holzbaur E, Tien M (1988) Structure and regulation of a lignin peroxidase gene from *Phanerochaete chrysosporium*. Biochem Biophys Res Comm 155:626–633

Hotelling H (1953) New light on the correlation coefficient and its transforms. J R Stat Soc Ser B 15:193–232

Intapruk C, Higashimura N, Yamamoto K, Okada N, Shinmyo A, Takano M (1991) Nucleotide sequences of two genomic DNAs encoding peroxidase of *Arabidopsis thaliana*. Gene 98:237–241

Kaput J, Goltz S, Blobel G (1982) Nucleotide sequence of the yeast nuclear gene for cytochrome c peroxidase precursor: functional implications of the pre sequence for protein transport into mitochondria. J Biol Chem 257:15054–15058

Lagrimini L, Burkhart W, Moyer M, Rothstein S (1987) Molecular cloning of complementary DNA encoding the lignin-forming peroxidase from tobacco: Molecular analysis and tissue-specific expression. Proc Natl Acad Sci USA 84:7542–7546

Lesk A, Chothia C (1982) Evolution of proteins formed by β-sheets II. The core of the immunoglobulin domains. J Mol Biol 160:325–342

Magnusson R, Gestautas J, Seto P, Taurog A, Rapoport B (1986) Isolation and characterization of a cDNA clone for porcine thyroid peroxidase. FEBS Lett 208:391–396

Mazza G, Welinder K (1980) Covalent structure of turnip peroxidase 7. Cyanogen bromide fragments, complete structure, and comparison to horseradish peroxidase C. Eur J Biochem 108:481–489

Morgens P, Callahan A, Dunn L, Abeles F (1990) Isolation and sequencing of cDNA clones encoding ethylene-induced putative peroxidases from cucumber cotyledons. Plant Mol Biol 14:715–725

Morishita K, Tsuchiya M, Asano S, Kaziro Y, Nagata S (1987) Chromosomal gene structure of human myeloperoxidase and regulation of its expression by granulocyte colony-stimulating factor. J Biol Chem 262:15208–15213

Roberts E, Kolattukudy P (1989) Molecular cloning, nucleotide sequence, and abscisic acid induction of a suberization-associated highly anionic peroxidase. Mol Gen Genet 217:223–232

Roberts E, Kutchan T, Kolattukudy P (1988) Cloning and sequencing of cDNA for a highly anionic peroxidase from potato and the induction of its mRNA in suberising potato tubers and tomato fruits. Plant Mol Biol 11:15–31

Sellers P (1974) On the theory and computation of evolutionary distances. J Soc Industrial and Applied Mathematics (SIAM) 26:787–793

Smith A, Santama N, Dacey S, Edwards M, Bray R, Thorneley R, Burke J (1991) Expression of a synthetic gene for horseradish peroxidase C in *Escherichia coli* and folding and activation of the recombinant enzyme with Ca^{2+} and heme. J Biol Chem 265:13335–13343

Smith T, Waterman M, Fitch W (1981) Comparative biosequence metrics. J Mol Evol 18:38–46

Sneath P, Sokal R (1973) Numerical taxonomy. WH Freeman, New York

Sokal R, Rohlf F (1981) Biometry. WH Freeman, New York

Tyson H (1992) Relationships between amino acid sequences determined through optimum alignments, clustering and specific distances patterns; application to a group of scorpion toxins. Genome 35:360–371

Welinder K (1979) Amino acid sequence studies of horseradish peroxidase. Amino and carboxyl termini, cyanogen bromide and tryptic fragments, the complete sequence, and some structural characteristics of horseradish peroxidase C. Eur J Biochem 96:483–502

Wilkinson L (1989) SYSTAT: the system for statistics. Systat Inc

Yamada M, Hur SJ, Hashinaka K, Tsuneoka K, Saeki T, Nishio C, Sakiyama F, Tsunasawa S (1987) Isolation and characterization of a cDNA coding for human myeloperoxidase. Arch Biochem Biophys 255:147–155