Bona Fide Prediction of Aspects of Protein Conformation

Assigning Interior and Surface Residues from Patterns of Variation and Conservation in Homologous Protein Sequences

Steven A. Benner, Ian Badcoe†, Mark A. Cohen and Dietlind L. Gerloff

Laboratory for Organic Chemistry
E.T.H. 8092 Zurich, Switzerland

Heuristics have been developed for analyzing patterns of conservation and variation within a set of aligned homologous protein sequences for the purpose of assigning amino acids whose side-chains lie on the surface and inside the folded structure of a protein. These were used in several recent *bona fide* predictions of the secondary structure of proteins from sequence data, made and published before crystallographic information became available. Heuristics based on concurrent hydrophilic variation identify positions that lie on the surface. Heuristics based on concurrent hydrophobic conservation and variation identify positions lying in the interior. These heuristics are described here in detail and their performance evaluated when applied to seven protein families with known three-dimensional structures. The performance of individual heuristics is shown to depend on the nature of the multiple alignment within the protein family, and a strategy is presented for obtaining surface and interior assignments useful for predicting secondary structure.

Keywords: protein structure prediction; surface accessibility; evolution; neutral mutation

1. Introduction

Two complementary challenges presently define the frontier of structural biology in proteins: design and prediction. The design challenge will be met when biological chemists routinely invent polypeptides that fold and catalyze reactions. The prediction challenge will be met when biological chemists routinely predict the conformation of polypeptide sequences created by evolutionary processes to fold and catalyze reactions.

Improvements in methods for synthesizing and purifying polypeptides have enabled steady progress towards the first goal (Sheehan *et al.*, 1966; Chakravaty *et al.*, 1973; Gutte *et al.*, 1979; Allemann, 1989; Johnsson *et al.*, 1990; Hahn *et al.*, 1990). In one case, the solution structure of a designed enzyme has been proven by multidimensional NMR and its catalytic mechanism explored by physical organic methods (Johnsson *et al.*, 1990). The simpler goal, designing a polypeptide that folds, has been approached in still more laboratories (Kaiser, 1988; Eisenberg *et al.*, 1986; Hecht *et al.*, 1990; Goraj *et al.*, 1990; Padmanabham *et al.*, 1990), and the conformation of an additional designed protein has now been established by NMR (Osterhout *et al.*, 1992).

Similar progress is now being made towards meeting the prediction challenge (Fasman, 1989; Overington *et al.*, 1990). This progress is demonstrated most recently by *bona fide* predictions, those made and published before crystallographic or NMR data are available, of the conformation of several proteins (Crawford *et al.*, 1987; Bazan, 1990; Benner & Gerloff, 1991; Russell *et al.*, 1992; Musacchio *et al.*, 1992*a*; Benner *et al.*, 1993*a*,*b*; Gerloff *et al.*, 1993*a*,*b*) and later analyzed using subsequently determined structures (Hyde *et al.*, 1988; Knighton *et al.*, 1991; de Vos *et al.*, 1992; Musacchio *et al.*, 1992*b*; Yu *et al.*, 1992; Kim & Rees, 1992; Waksman *et al.*, 1992). These predictions join others made with the help of spectroscopic information that provided conformational clues. Three key examples are the bacterial aspartate receptor (Moe & Koshland, 1986; Milburn *et al.*, 1991), interleukin II (Cohen *et al.*, 1986; Cohen & Kuntz, 1987) and interleukin IV (Curtis *et al.*, 1991). In at least one case, a prediction has evidently been a more accurate representation of reality than a published crystal structure (Brandhuber *et al.*, 1987; Bazan, 1992; McKay, 1992).

† Present address: Department of Biochemistry, School of Medical Science, University of Bristol, Bristol, BS 1TD, England.

0022-2836/94/030926-33 $08.00/0

Bona fide prediction tests are exacting. They expose both the prediction method and its user to the risk of public failure. They therefore force clear decisions concerning prediction strategy. Further, they exclude all possible biases that could arise from knowledge of the conformation of the target protein. In particular, the solved structure cannot be used to parameterize the computerized prediction methods, cannot influence a conformational model assembled *via* human intervention, and cannot help the predictor select which predictions to disclose. Such factors are well known in chemistry to bias the output of highly parameterized computational methods even when steps are taken to exclude this bias (for one of many documented cases, see Wentrup, 1984).

Many of the *bona fide* predictions of secondary structure made to date have followed a research approach outlined over a decade ago by Lenstra *et al.* (1977), Garnier *et al.* (1978), Maxfield & Scheraga (1979) and others. In this approach, classical methods are used to predict secondary structure for individual members of a family of homologous protein sequences. These predictions are then averaged over all family members to yield a consensus prediction. This approach rests on the assumption that the folded structures of homologous proteins are similar (Chothia & Lesk, 1986; Summers *et al.*, 1987; Blundell *et al.*, 1987; Bowie *et al.*, 1991), and assumes that errors in individual classical predictions are distributed randomly about reality.

Kirschner and his co-workers pioneered the application of this approach in *bona fide* structure prediction. They averaged the secondary structure predictions made by applying the standard GOR heuristic (Garnier *et al.*, 1978) to each of ten homologous sequences of the alpha subunit of tryptophan synthase. The resulting average prediction showed approximately eight beta strands interspersed by eight alpha helices. Crawford *et al.* (1987) recognized this pattern as an indicator of a particular type of fold: the 8-fold alpha-beta barrel (Farber & Petsko, 1990). A subsequently determined crystal structure showed that their inferences were largely correct (Hyde *et al.*, 1988).

This approach does not, unfortunately, appear to be generally useful (Lenstra *et al.*, 1977). Normally, the consensus predictions have only marginally improved three state (alpha helix, beta strand, or neither alpha helix nor beta strand) per residue scores (Zvelebil *et al.*, 1987). Nor does the approach improve predictions as evaluated by other, more meaningful, scoring methods.

Nevertheless, the approach fits well one widely accepted research paradigm in the contemporary field of structure prediction. This paradigm has as its goal the development of an automated process that accepts one or more polypeptide sequences as input and yields a secondary structure prediction as output, without considering tertiary structure. In the paradigm, the success of the process is measured by automatic tabulation of three state per-residue scores over a statistically large number of proteins (Robson & Garnier, 1993). Classical versions of this paradigm include prediction programs based on Chou-Fasman (Chou & Fasman, 1978) and GOR (Garnier *et al.*, 1978) heuristics found in standard packages of sequence analysis software. Other approaches for achieving this goal include neural networks applied to homologous protein sequences made available to the public by server (Rost & Sander, 1992).

One can attempt to evaluate this paradigm in light of the experience of organic chemistry generally, where efforts to understand how constitution determines conformation have a long history. Some time ago (Benner, 1989), we noted that chemical experience suggested that this paradigm was not likely to be broadly successful. In particular, distinctions between local conformation (e.g. secondary structure) and global conformation (e.g. tertiary structure) are arbitrary and interdependent, often frustrating efforts to analyze conformation, even in molecules much smaller than proteins. It is rarely productive in chemistry to treat molecular behavior as a statistical average over many different molecules. Further, automation disconnects the chemist from involvement with the very chemical details that are necessary to develop understanding. Thus, no conformational problem in chemistry has yet been solved by an initial focus on developing an automated computer program.

Instead, chemistry uses a quite different research paradigm to solve problems in conformational analysis. In this paradigm, each molecule is analyzed individually to develop an understanding of underlying principles in the individual case. This understanding is then tested by applying it to another individual case, and then to another. Errors offer insights that suggest modifications and improvements in the formalism. Gradually, a formalism based on underlying structural principles is built, together with an expertise that resides within the chemist. Automation is the final step in the process, appropriate only when understanding is firmly in place.

This procedure requires, of course, involvement of a chemist with a certain degree of expertise. Several have criticized the approach for this reason, arguing that automated systems are intrinsically superior (Robson & Garnier, 1993; Rost & Sander, 1992; Rost *et al.*, 1993). In fact, an approach that involves interaction by a chemist is always superior to an automated approach before a problem is understood.

Over the past seven years, we have been building a chemical formalism for extracting conformational information from a set of aligned homologous protein sequences by making *bona fide* predictions for specific protein families. These have been published as worked examples (Benner, 1989; Benner & Gerloff, 1991; Benner *et al.*, 1992; Benner, 1992*b*; Benner *et al.*, 1993*a*; Gerloff *et al.*, 1993*a*,*b*), providing the biochemist with access to the expertise needed to solve his/her own problems. As a result of this work, a set of effective tools has been developed for predicting secondary structure and certain elements of tertiary structure from a set of aligned homologous protein sequences. For example, in the prediction for collagenase, a protein with just over 200 amino acids, only 2 of the 130 residue assignments of secondary structure evidently misassigned an alpha helix for a beta strand or *vice versa*. The three state per residue score was evidently in excess of 70%, the upper limit of what can be obtained given current technology for assigning secondary structure to experimental data (Colloc'h *et al.*, 1993).

These results suggest that our level of understanding is adequate to attempt to automate the secondary structure prediction tools developed in Zürich, to test them "blind" against a set of proteins with known three-dimensional structures, and to attempt to build a still broader quantitative understanding of their effectiveness. This is the first of three papers reporting such analyses. This paper focuses on assigning surface and interior positions in a set of aligned homologous protein sequences. These assignments provide a form of tertiary structural information, essential for the assignment of secondary structure, which is the focus of the second paper. The third will focus on tools for assembling predicted secondary structural units into supersecondary and tertiary structural models.

2. Materials and Methods

(a) *Study objects*

Seven families of proteins were used in this study: aspartate aminotransferase (AAT), alcohol dehydrogenase (ADH), lactate dehydrogenase (LDH), myoglobin (MYO), phospholipase A (PLA), plastocyanin (PLC) and Cu/Zn superoxide dismutase (SOD). These families were chosen to meet 2 conditions. First, they contain an adequate number and evolutionary distribution of homologous sequences for the heuristics described here to be conveniently implemented. Second, a crystal structure of reasonable quality is available for 1 or more members of the family. The 7 protein families also represent some of the structural and mechanistic diversity found in natural proteins, including monomeric and multimeric species, metal-containing and metal-free proteins, and catalytic and binding proteins.

(b) *Crystal structure data*

Coordinates from crystal structures for a representative member of each protein family were obtained from the Brookhaven Data File (Bernstein *et al.*, 1977): aspartate aminotransferase (AAT) from *Escherichia coli* (2·8 Å resolution, Smith *et al.*, 1986), alcohol dehydrogenase (ADH) from horse liver (2·4 Å resolution, Eklund *et al.*, 1976), lactate dehydrogenase (LDH) from dogfish (2·0 Å resolution, White *et al.*, 1976), myoglobin (MYO) from sperm whale (1·4 Å resolution, Phillips, 1980), phospholipase A (PLA) from bovine pancreas (1·7 Å resolution, Dijkstra *et al.*, 1981), plastocyanin (PLC) from poplar (1·6 Å resolution, Guss & Freeman, 1983), and Cu-Zn superoxide dismutase (SOD) from bovine erythrocyte (2·0 Å resolution, Tainer *et al.*, 1982).

(c) *Alignments and evolutionary trees*

The DARWIN system (Gonnet & Benner, 1991) was used to create alignments and evolutionary trees. These were used unrefined. Alignments were constructed using the optimal mutation matrix and a linear approximation of the optimal gap scoring equations obtained from the exhaustive matching of the MIPS (version 64) protein sequence database (Gonnet *et al.*, 1992).

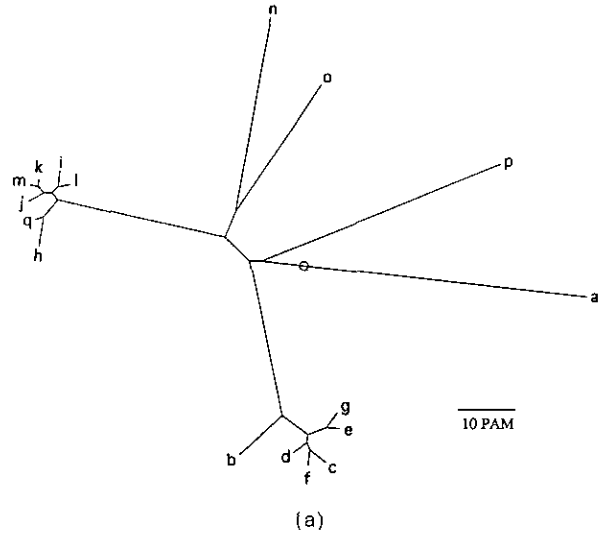
For each protein family, a set of pairwise alignments was constructed relating each family member with every other. An evolutionary distance (with a variance) was assigned for each pair. Evolutionary distance was measured in PAM (accepted point mutations per 100 aligned positions) units (Dayhoff *et al.*, 1978). This is more precise than pairwise identity, used previously as an estimate of evolutionary distance (Benner, 1989; Benner & Gerloff, 1991). The PAM distance separating 2 protein sequences may be understood in terms of a 1% PAM mutation matrix, a matrix describing the probability of matching every amino acid with every other amino acid in an alignment of 2 proteins divergent by 1 accepted point mutation per 100 residues (Dayhoff *et al.*, 1978). The PAM distance between 2 aligned protein sequences is the number of times the first protein sequence must be transformed using the 1% PAM matrix to achieve the second sequence with highest probability.

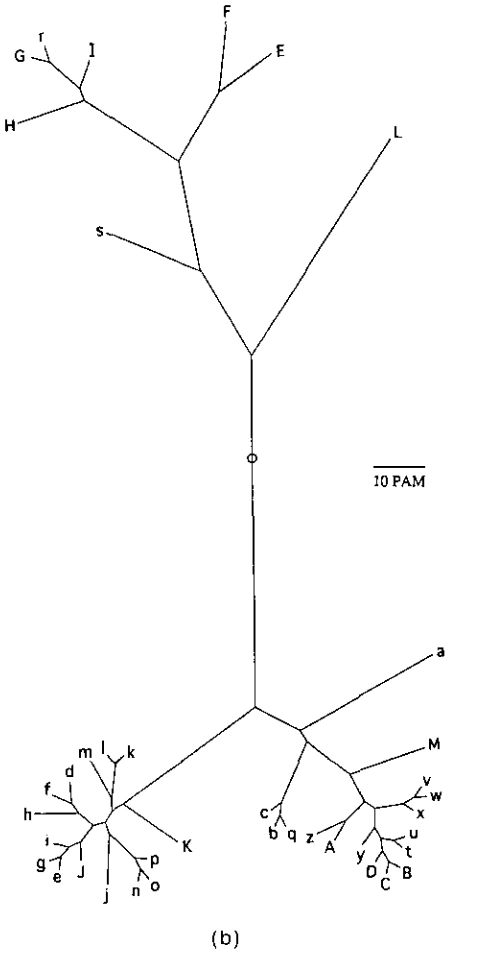
The connectivity of an evolutionary tree for each protein family was constructed from a PAM distance matrix for the constituent proteins. These are shown in Fig. 1. The lengths of lines between nodes within this tree was calculated by least-squares fit of the PAM distances, and probabilistic sequences for ancestral sequences represented by the internal nodes calculated as described elsewhere. Multiple alignments (available on request) were

Table 1
Description of protein families

Protein family[a]	PAM width	Number of proteins	Number of positions	Amino acids in crystal structure	Number on surface (%)	Number inside	Quaternary structure	Disulfide bonds?
AAT	104	17	416	396	213 (54)	183	Dimer	no
ADH	190	40	394	374	170 (45)	204	Dimer, tetramer	no
LDH	135	21	321	329	172 (52)	157	Tetramer	no
MYO	190	78	156	153	95 (63)	58	Monomer	no
PLA	160	78	136	120	69 (58)	51	Monomer	yes
PLC	102	28	102	99	64 (64)	35	Monomer	no
SOD	163	31	175	149	81 (54)	68	Dimer	no

[a] In all tables: AAT, aspartate aminotransferase; ADH, alcohol dehydrogenase; LDH, lactate dehydrogenase; MYO, myoglobin; PLA, phospholipase; PLC, plastocyanin; SOD, Cu/Zn superoxide dismutase. A surface residue is defined as one where the side-chain atoms are >50% accessible to a 1·4 Å probe in the reference crystal structure, except for Gly, where all atoms are considered.





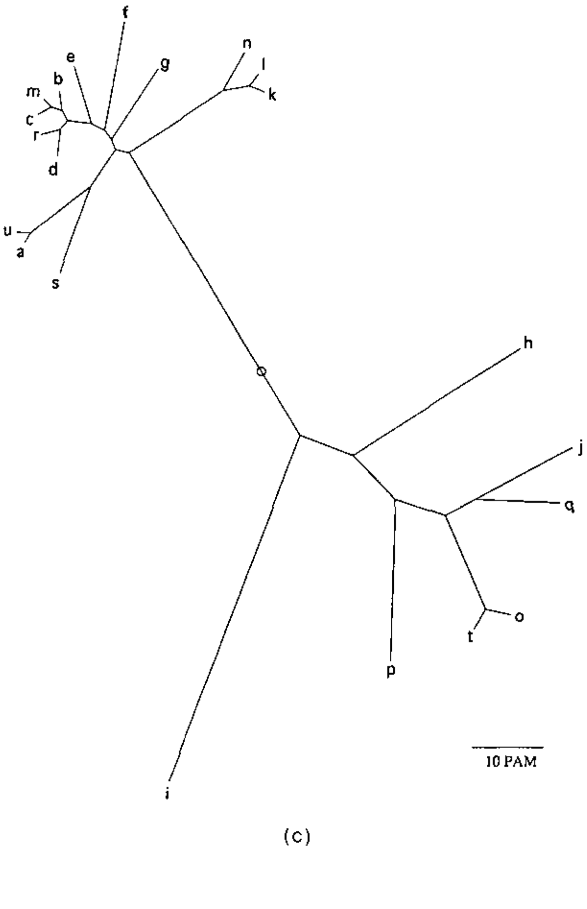
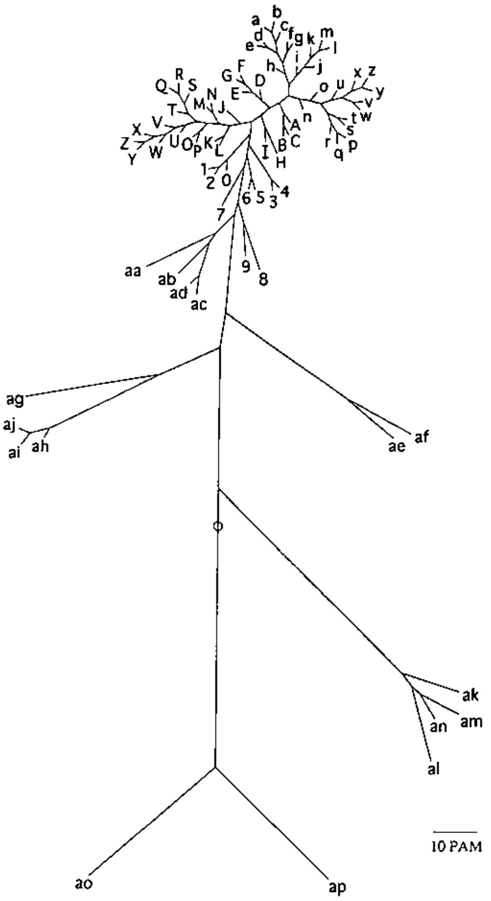
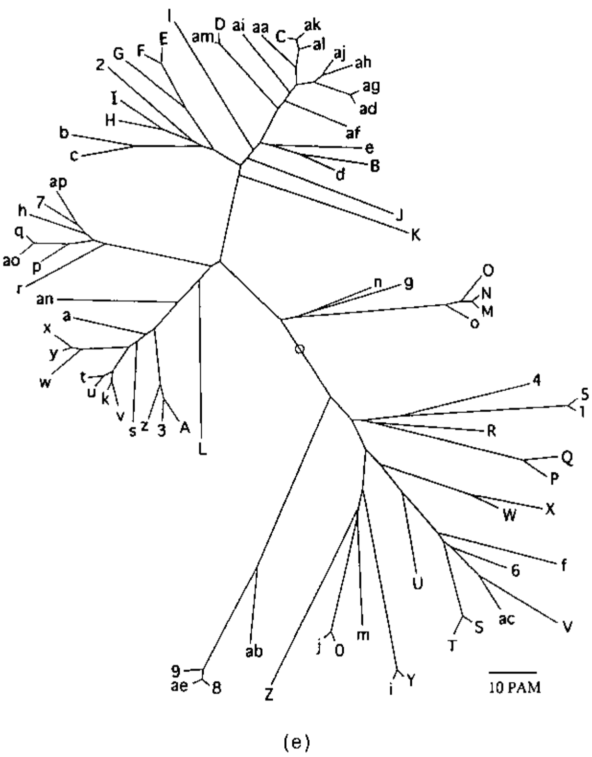


Figure 1. Unrooted phylogenetic trees of the 7 protein families used in this work: (a) aspartate aminotransferase (AAT); (b) alcohol dehydrogenase (ADH); (c) lactate dehydrogenase (LDH); (d) myoglobin (MYO); (e) phospholipase (PLA); (f) plastocyanin (PLC) and (g) Cu/Zn superoxide dismutase (SOD). The key is below; asterisks indicate which crystal structures were used to evaluate the heuristics. The trees were displayed and manipulated using the PhylTree package running under the Oberon operating system. The distances along the branches are proportional to the PAM distance as indicated by the bar in each case. The open circle represents the center of gravity of the trees and can be considered as the top-most node purely for convenience; it does not indicate a root. (a) Aspartate aminotransferase: a*, *E. coli*; b*, chicken, cytoplasmic; c, horse, cytoplasmic; d, human, cytoplasmic; e, mouse, cytoplasmic; f, pig, cytoplasmic; g, rat, cytoplasmic; h, chicken, mitochondrial; i, horse, mitochondrial; j, human, mitochondrial; k, mouse, mitochondrial; l, pig, mitochondrial; m, rat, mitochondrial; n, lupin; o, alfalfa, leaf; p, yeast (*S. cerevisiae*), cytosolic; q, turkey, mitochondrial. (b) Alcohol dehydrogenase: a, human, adh pi chain; b, human class III adh chi subunit; c, rat, adh 2; d, maize, adh 2; e, maize, adh 1; f, barley, adh 3; g, pearl millet, adh; h, barley, adh 2; i, barley, adh 1; j, strawberry, adh; k, garden pea, adh; l, white clover, adh 1; m, mouse-eared cress, adh; n, potato, adh 2; o, potato, adh 3; p, potato, adh 1; q, horse, adh III chi chain; r, *S. cerevisiae*, adh 1; s, *S. pombe*, adh 1; t, human, adh gamma-1 chain; u, human, adh gamma-2 chain; v, horse, adh S chain; w*, horse, adh E chain; x, chicken, adh; y, human, adh alpha-chain; z, rat, adh 1; A, mouse, adh alpha-chain; B, human, adh beta-2; C, human, adh beta-1; D, baboon, adh; E, *E. nidulans*, adh 1; F, *E. nidulans*, adh 3; G, *S. cerevisiae*, adh 2 H, *S. cerevisiae*, adh 3; I, *K. lactis*, adh 1; J, rice, adh 1; K, rice, adh 2; L, *Z. mobilis*, adh 1; M, Japanese quail, adh. (c) Lactate dehydrogenase: a, chicken, ldh-H; b, Human, ldh-M; c, mouse, ldh-M; d, pig, ldh-M; e, chicken, ldh-M; f*, spiny dogfish, ldh-M; g, human, ldh-X; h, *Lactobacillus casei*, ldh; i, *Thermus aquaticus*, ldh; j, *Bacillus subtilis*, ldh; k, mouse, ldh-M; l, mouse, ldh-X; m, rat, ldh-M; n, rat, ldh-X; o, *Bacillus caldotenax*; p, *Bacillus megaterium*; q, *B. psychrosaccharolyticus*; r, rabbit, ldh-M; s, *F. heteroclitus*, ldh-H; t, *B. stearothermophilus*, ldh; u, duck, ldh-B. (d) Myoglobin: a, *G. gorilla beringei* (gorilla); b, *H. sapiens* (human); c, *P. troglodytes* (chimpanzee); d, *H. agiles* (agile gibbon); e, *P. pygmaeus* (orangutang); f,

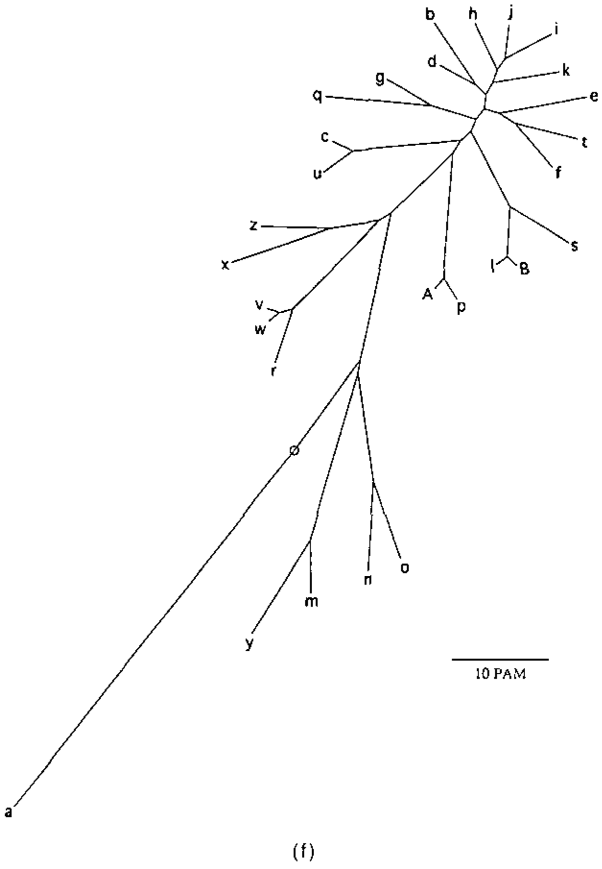

a b c d e f g k m l i j h F G D E Q R S N M J T X V Z Y W U O P K L 1 2 0 7 6 5 3 4 I B C H A n o u x z y v t w s r q p 9 8 aa ab ad ac ag aj ai ah af ae ak am an al ao ap
10 PAM

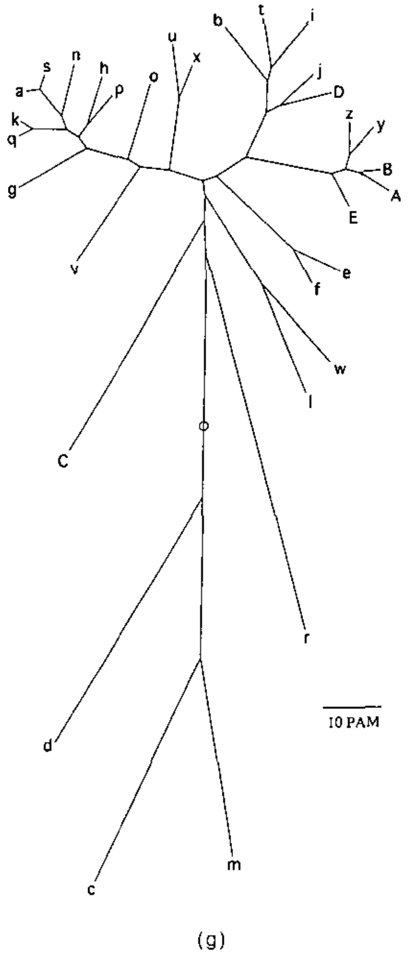
(d)


l E D ai aa C ak am al aj ah ag ad af e B d J K 2 G F I H b c ap h 7 q ao p r an a x y w t u k v s z 3 A L n g O N M o 4 5 1 R Q P X W f 6 U ac V S T ab m j o i Y 9 ae 8 Z
10 PAM

(e)

M. fascicularis (crab eating macaque); g, *P. anubis* (olive baboon); h, *C. fiber* (beaver); i, *L. lagothricha* (common wooly monkey); j, *S. sicureus* (common squirrel monkey); k, *C. apella* (brown-capped capuchin); l, *A. trivirgatus* (night monkey); m, *C. jacchaus* (common marmoset); n, *O. cuniculus* (rabbit); o, *O. afer* (aardvark); p, *P. guairae* (casiragua); q, *L. maximus* (plains viscacha); r, *C. gundi* (gundi); s, *E. europaeus* (Western European hedgehog); t, *R. aegyptiacus* (Egyptian rousette); u, *S. leucodon Ehrenbergi* (Ehrenberg's mole rat); v, *L. lutra* (European river otter); w, *M. meles* (Eurasian badger); x, *O. princeps* (Southern American pika); y, *T. glis* (tree shrew); z, *S. scrofa* (pig); A, *G. crassicaudatus* (thick-tailed galago); B, *N. coucang* (slow loris); C, *P. potto Edwarsi* (potto); D, *Z. californianus* (Californian sea lion); E, *V. chama* (Cape fox); F, *C. familiaris* (dog); G, *L. pictus* (African hunting dog); H, *M. rufus* (red kangaroo); I, *D. marsupialis Virginiana* (N. American opossum); J, *H. grypus* (gray seal); K, *L. mustelinus* (weasel lemur); L, *E. caballus* (horse); M, *Z. cavirostris* (goose-beaked whale); N, *M. carlhubbsi* (Hubb's whale); O*, *P. catodon* (sperm whale); P, *K. simus* (dwarf sperm whale); Q, *E. gibbosus* (California gray whale); R, *M. novaeangliae* (humpback whale); S, *B. acutorostrata* (Minke whale); T, *B. ahysalus* (finback whale); U, *I. geoffrensis* (Amazon dolphin); V, *P. phocoena* (harbor porpoise); W, *D. delphis* (saddleback dolphin); X, *T. truncatus* (Atlantic bottle-nosed dolphin); Y, *G. melaena* (long-finned pilot whale); Z, *O. orca* (killer whale); 0, *B. taurus* (cow); 1, *C. elaphus* (red deer); 2, *O. aries* (sheep); 3, *E. maximus* (Indian elephant); 4, *L. africana* (African elephant); 5, *O. anatinus* (duckbill platypus); 6, *T. aculeatus Aculeatus* (Australian echidna); 7, *M. musculus* (mouse); 8, *A. forsteri* (emperor penguin); 9, *G. gallus* (chicken); aa, *A. mississippiensis* (American alligator); ab, *V. varius* (lace monitor); ac, *C. mydas Caranigra* (green sea-turtle); ad, *G. geographica* (map turtle); ae, *T. albacares* (yellowfin tuna); af, *C. carpio* (common carp); ag, *H. portusjacksoni* (Port Jackson shark); ah, *M. antarcticus* (gummy shark); ai, *G. australis* (shark); aj, *G. japonicus* (shark); ak, *D. auricularia* (sea hare); al, *A. juliana* (sea hare); am, *A. kurodai* (Kuroda's sea hare); an, *A. limacina* (slug sea hare); ao, *B. canaliculatum* (channeled whelk); ap, *C. rhizophorarum* (water snail). (e) Phospholipase A_2: a, Shield snake; b, Yellow-lipped sea krait; c, Yellow-lipped sea krait, inactive; d, Common tiger snake; e, Olive sea snake; f, Tropical rattle snake, crotoxin B; g, Banded krait, basic PLA; h, Rat gastric/lung; i, Tropical rattle snake, crotoxin A, inactive; j, Western sand viper, ammodytoxin C; k, Indian cobra; l, Common tiger snake, PLA homolog; m, Bothrops asper; n, Many-banded krait, non-toxic isoform; o, Many-banded krait, isoform A4; p*, Pig pancreatic; q*, Bovine pancreatic; r, Equine pancreatic; s, Ringhals; t, a) Monocled cobra II *N.N. siamensis*. u, b) Monocled cobra III *N.N. siamensis*. v, Chinese cobra *N.N. atra*, not toxic. w, Forest cobra I *N. melanoleuca*. x, a) Forest cobra II *N. melanoleuca*. y, b) Forest cobra III *N. melanoleuca*. z, Mozambique cobra I *N. mossambica*. A, Mozambique cobra 3 *N. mossambica*. B, Common tiger snake notechis1 homolog; C, Mulga snake Pa11; D, Mulga snake Pa13; E, Common tiger snake notexin; F, Common tiger snake notechis5; G, Beaked sea snake; H, Broad-banded sea krait I; I, Broad-banded sea krait III; J, Australian taipan taipoxin alpha; K, Australian taipan taipoxin beta; L, Australian taipan taipoxin gamma prec. M, beta bungarotoxin A1, M. banded krait; N, beta bungarotoxin A_2, M. banded krait; O, beta bungarotoxin A3, M. banded krait; P, Gabon viper; Q, Rhinoceros viper; R, Horned viper, caudoxin; S, Eastern diamond backed rattlesnake; T*, Western diamond backed rattlesnake; U, Habu snake I, *T. flavoviridis*. V, Himehabu, *T. okinavensis*. W,





Mamushi, *A. blomhoffi*. X, Habu snake X, *T. flavoviridis*. Y, Tropical rattle snake, crotoxin basic; Z, Eastern cottonmouth; 0, Western sand viper (A), ammodytoxin A; 1, Western sand viper (A), inhibitor; 2, Blue-ringed sea krait; 3, Spitting cobra, nigexine; 4, Western sand viper (B); 5, Western sand viper (B) inhibitor; 6, Halys viper acidic PLA; 7, Human pancreatic; 8, Rat platelet (version1); 9, Rat platelet (version2); aa, Red-bellied black snake, pseudexin A; ab, Human, sinovial fluid; ac, Halys viper; ad, Mulga snake Pa 1G; ae, Rat II; af, Red-bellied black snake, pseudexin B; ag, Mulga snake Pa 3; ah, Mulga snake Pa 5; ai, Mulga snake Pa 9C; aj, Mulga snake Pa 10A; ak, Mulga snake Pa 12A; al, Mulga snake Pa 12C; am, Mulga snake Pa 15; an, Banded krait, neutral PLA; ao, Sheep, pancreatic; ap, Canine. (f) Plastocyanin: a, *Anabaena variabilis*. b, kidney bean; c, broad bean; d, garden lettuce; e, European elder; f, vegetable marrow; g, shepherd's purse; h, dog's mercury; i, potato; j, Chilean potato-tree; k, bitter dock; l*, Lombardy poplar; m, *Chlorella fusca*. n, *Enteromorpha prolifera*. o, sea lettuce; p, white campion; q, mouse-ear cress; r, barley; s, Lombardy poplar b; t, spinach; u, garden pea; v, rice; w, rice; x, carrot; y, *Scenedesmus obliquus*. z, parsley; A, *Silene pratensis*. B, popni. (g) Cu/Zn superoxide dismutase: a*, cow; b, cabbage; c, *B. abortus*. d, *C. crescenus*. e, *D. melanogaster*. f, *D. virilis*. g, horse; h, human; i, tomato; j, maize 2; k, mouse; l, *N. crassa*. m, *P. leiognathi*. n, pig; o, blue shark; p, rabbit; q, rat; r, *S. mansoni*. s, sheep; t, spinach; u, *X. laevis* 1. v, swordfish. w, *S. cerevisiae*. x, *X. laevis* 2. y, tomato (chloroplast); z, garden pea (chloroplast); A, petunia (chloroplast); B, spinach (chloroplast); C, human (extracellular); D, Scots pine 1; E, Scots pine 2

Fig. 1. (*continued*)

built in this process. These alignments were used directly without adjustment (which would be part of a refinement procedure). In particular, the alignment was not adjusted using crystallographic information, as this would not be possible in a *bona fide* structure prediction situation. The single-letter code for amino acids is used throughout.

A family of proteins can be divided into subfamilies defined by the maximum PAM width (MaxPW). This is the PAM value assigned to the highest bridge connecting proteins within the subfamily. The higher the PAM value of this bridge, the more sequence divergence the proteins in the subfamily display overall.

A summary of various parameters of the alignment of the families of proteins examined is shown in Table 1.

(d) *Surface accessibility*

Solvent-accessible surfaces were calculated using the program of Connolly (1983*a*,*b*). This program probes the structure with a sphere of specified size, yielding as output a set of points occupied by the center of the probe as it rolls over the surface of the protein. The area associated with each point is also calculated. In this study, the probe radius was 1·4 Å and points were generated at a density of

Table 2
Surface exposures in the test protein families

Protein	>30% Exposed		>40% Exposed		>50% Exposed		>60% Exposed		>70% Exposed	
family*	#	(%)	#	(%)	#	(%)	#	(%)	#	(%)
AAT	271	65·1	238	57·2	213	51·2	180	43·3	143	34·4
ADH	228	57·9	198	50·3	170	43·1	144	36·5	113	28·7
LDH	220	68·5	193	60·1	170	53·0	152	47·4	123	38·3
MYO	112	71·8	103	66·0	94	60·3	80	51·3	60	38·5
PLA	88	64·7	78	57·4	69	50·7	63	46·3	55	40·4
PLC	70	68·6	65	63·7	62	60·8	50	49·0	33	32·4
SOD	101	62·0	91	55·8	81	49·7	67	41·1	51	31·3

A surface residue is defined by a side-chain accessibility to a 1·4 Å probe in the reference crystal structure, except for Gly, where all atoms are considered.

10 Å^{-2}. The program was used in its original form, except that the accuracy of output of areas was increased to 7 decimal places. The areas associated with the surface points contacting each atom were then summed. Except for glycine (see below), the main-chain atoms (C^{α}, N, C and O) were not included in this summation. Where a point contacted several atoms, its area was divided equally between them. The exposed surface areas of individual residues were calculated by summing the exposed areas of each atom.

Side-chain exposures were defined as a percentage of the maximal exposure for the specific residue type. Maximal exposure was defined as the exposure of each residue side-chain in an extended conformation (all torsion angles 180°) in a peptide flanked by 2 glycine residues. Using these procedures for a typical protein, only approx. 1% of residues show exposures in excess of 100%. In the case of glycine, surface exposure is expressed as a fraction of total possible exposure of backbone atoms. A summary of the surface exposures in the 7 protein families studied in this work is shown in Table 2.

(e) *Evaluating heuristics*

Most heuristics discussed here make binary assignments (surface or interior). These must be evaluated in light of surface-accessibility parameters, which display a continuum of exposure of amino acid side-chains. A useful cutoff to divide this continuum to define interior and surface is taken to be one that would allow correct assignment of surface alpha helices in their patterns (Benner, 1989; Benner & Gerloff, 1991; Benner *et al.*, 1992; Benner *et al.*, 1993*a*; Gerloff *et al.*, 1993*a*,*b*). For each surface helix in the 7 test structures, helical wheels were constructed with "inside" and "outside" positions assigned according to different definitions of surface. The number of "perfectly amphiphilic" helices, "amphiphilic helices missing ends" and "predictable helices" were counted (Table 3), as were the number of helices that were not predicted because no 3·6 residue periodicity could be detected. Exposure of side-chains was typically greater at the ends of the helices, making more stringent definitions of surface more useful at the ends of helical segments. Nevertheless, cutoffs at 30% and 70% were not useful (Table 3), while surface definitions of >50% exposure gave the highest number of correct helix assignments. Therefore, >50% side-chain exposed was chosen to define a surface residue for the purpose of scoring heuristics. Any residue whose side-chain is less than 50% exposed was considered to be interior.

Heuristics for assigning surface and interior positions were evaluated using 2 scores: accuracy and coverage. Accuracy is defined by the ratio of the number of correctly assigned positions divided by the total number of assignments made. For example, the accuracy of a surface assignment is defined as (surface assignments that are correct)/(total surface assignments). Coverage is defined as the ratio of the number of correct assignments of a particular class divided by the total number of positions in that class. For example, surface coverage = (assignments to surface positions)/(total number of surface positions). Each score is expressed as a percentage; the higher the value of each of the 2 scores, the more successful the heuristic.

(f) *Implementing heuristics*

The work performed here was done using programs written in Fortran and compiled and run on a Sun SparcStation 2 using a Unix operating system.

3. Results

(a) *Theory*

Four generalizations apply to protein structure (Schulz & Schirmer, 1979).

(1) Hydrophobic residues tend to lie inside the folded structure.

(2) Hydrophilic residues tend to be on the surface.

(3) Conserved residues tend to lie inside, or near the active site.

(4) Variable residues tend to lie on the surface of proteins (Hubbard & Blundell, 1987; Lim & Sauer, 1989).

These generalizations apply in a majority of instances (Shrake & Rupley, 1973). However, they are far from universal (Lee & Richards, 1971), and the lack of universality has precluded their use as the basis for reliable structure predictions, at least in their simplest form. For example, methods that attempt to assign secondary structure by patterns of amphiphilicity in polypeptide segments yield good, but not excellent, assignments (Lim, 1974*a*,*b*). Similarly, efforts to predict surface residues by their variability and active-site residues based on conserved functionality yield good, but not excellent, predictions (Zvelebil *et al.*, 1987; Zvelebil & Sternberg, 1988). In most proteins, at least some hydrophobic residues lie on the surface of the folded structure, at least some uncompensated hydrogen-

Table 3

Determining the definition for surface for testing heuristics for predicting alpha helices

Protein family[a] (assignments)[b]	Number of helices in structure	Accesibility criterion for surface (%)	Perfectly amphiphilic helices[c]	Amphiphilic helices missing ends[d]	Predictable helices[e]	Total predictable (%)	Helices not found[f]
AAT (SwissProt)	12	70	2	···	4	6 (50)	6
		60	4	1	5	10 (83)	2
		50	7	1	4	12 (100)	0
		40	4	1	7	12 (100)	0
		30	4	1	5	9 (75)	2
ADH (PDB)	8	70	3		1	4 (50)	4
		60	2	1	3	6 (75)	2
		50	4	1	1	6 (75)	2
		40	4	1	1	6 (75)	2
		30	1	1	4	6 (75)	2
LDH (PDB)	11	70	1	1	6	8 (73)	3
		60	1	1	6	8 (73)	3
		50	2		6	8 (73)	3
		40	1	1	5	7 (64)	4
		30	1	—	3	4 (36)	7
MYO (PDB)	8	70	2	—	5	7 (88)	1
		60	2	—	5	7 (88)	1
		50	1		5	6 (75)	2
		40			4	4 (50)	4
		30			1	1 (13)	7
PLA (SwissProt)	3	70			2	2 (67)	1
		60	—	—	3	3 (100)	0
		50	—	—	3	3 (100)	0
		40	1	—	2	2 (67)	1
		30	—	—	1	1 (33)	2
Total	42	70	8	1	18	27 (64)	15
		60	9	3	22	34 (81)	8
		50	14	2	19	35 (83)	7
		40	10	3	19	32 (76)	10
		30	6	2	14	22 (52)	20

[a] SOD and PLC have no helices.

[b] SwissProt definition of helices are used for AAT and PLA. Helices defined by the SwissProt regimen are typically shorter than those defined by the PDB files, and very short helices are generally ignored.

[c] Perfectly amphiphilic helices meet the following criteria: (1) A fraction of 50% (±1 residue; minimum: 3 residues) of the segment's positions appears as surface on the same half-arc of the helical plot; (2) there is no disruption of the surface-arc by a position with solvent-accessibility up to 10% (strong interior criterion), nor disruption of an interior arc by a residue with a surface accessibility greater than 90%, is found; (3) the fraction of additional surface assignments appearing on the non-surface half-arc does not exceed 10% of the segment's positions.

[d] Amphiphilic helices missing ends meet the following criterion: the requirements for a perfectly amphiphilic helix can be achieved by truncating the helical segment at one or both ends by no more than 4 residues. Note: the definition of the ends of helical segments is not well-established, even given a high resolution crystal structure.

[e] Predictable helices meet the following criteria: (1) At least 3 surface assignments appear on the same half-arc of the helical plot; (2) no more than 1 disruption of the surface-arc by a position having a solvent accessibility parameter less than 10% (strong interior criterion), or the interior arc by a residue with a surface accessibility greater than 90%, is found; (3) the fraction of additional surface assignments on the non-surface half-arc is smaller than 20% of the segment's positions.

[f] Helices not found meet the requirements for none of the amphiphilicity classes outlined above.

bonding side-chains lie inside a structure, and at least some variation occurs near the active site. These and other violations of the simple structural "rules" generally frustrate efforts to predict the conformation of a polypeptide chain from a single sequence.

To obtain useful prediction tools from these generalizations, two factors must be considered. First, abundant evidence suggests that extreme conformational stability is possible in a folded protein structure if all available stabilizing interactions are exploited. Such extreme stability is, however, undesired in a protein that under typical physiological conditions, must be degraded and recycled. Thus, natural proteins have evolved to violate folding "rules" to engineer instability into the folded structure (Benner, 1989; Benner & Ellington, 1990). This hypothesis has two consequences of importance: (1) the conformational stability of native proteins can readily be improved by point mutation, and (2) natural protein sequences are deceptive as a guide to folded structure, even to those who might fully understand the rules governing protein conformation.

Second, a meaningful analysis of sequence conservation and variation is possible only in the context of a quantitative understanding of the divergence separating the protein sequences being compared, and then only in the context of an evolutionary tree. Conservation is most significant within a set of proteins that have, overall, undergone substantial sequence divergence. Variation is most significant

Table 4

Scoring the simple surface heuristic. Dependence of coverage and accuracy scores on definition of surface residue

%	% of correct assignments (accuracy)								% surface found (coverage)							
Exposure	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average
70	65·9	82·5	71·0	62·7	69·2	70·4	71·1	70·4	20·3	41·6	39·8	61·7	81·8	57·6	62·7	52·2
60	81·8	89·5	81·2	65·7	73·8	88·9	82·2	80·4	20·0	35·4	36·8	63·7	76·2	48·0	55·2	47·9
50	93·1	91·2	85·5	94·9	78·5	100·0	97·8	91·6	19·2	30·6	34·7	59·6	73·9	43·5	54·3	45·1
40	93·1	94·7	88·4	100·0	87·7	100·0	97·8	94·5	17·2	27·3	31·6	57·3	73·1	41·5	48·4	42·3
30	95·5	98·2	91·3	100·0	92·3	100·0	97·8	96·4	15·5	24·6	28·6	52·7	68·2	38·6	43·6	38·8

The heuristic assigns a position in an alignment if it contains at least 2 subfamilies (defined at a maximum PAM width of 120) that both contain one of the amino acids KREND and contain more than one type of amino acid (i.e. is variable). This corresponds to distribution criterion A (see the text, and Fig. 3).

The assignments are scored with respect to their accuracy (percentage of the assignments that are correct) and their coverage (fraction of surface positions identified) by comparison with a crystal structure of a representative member of the indicated protein family. In the crystal structure, individual amino acids are assigned to the surface depending on the percentage of the side-chain exposed to a solvent probe, as described in the text. The percentage exposure used to define surface to evaluate the heuristic is shown in the left column. Averages are unweighted, and have no exact interpretation.

within a set of proteins that have undergone relatively little sequence divergence overall. Consensus sequences are best constructed *per stirpes*, where lineages derived from a common ancestor receive equal weight regardless of the number of representative sequences available from each lineage.

To illustrate how these considerations influence the generation of structure prediction tools, consider a simple heuristic suggested by generalizations (2) and (4) for identifying positions in a multiple alignment of residues whose side-chains lie on the surface of a folded structure. In this heuristic, positions in an alignment that are variable are assigned to the surface; conserved positions are not. Further, at least one amino acid at the variable position must be "surface indicating" to merit a surface assignment, where surface indicating amino acids are *defined (for this example) to be* Lys, Arg, *Glu*, Asp, Ser, Asn and Gln.

In its simplest form, this heuristic is clearly naive unless the evolutionary distance between the proteins where variation is sought is specified. Intuitively, one expects such a surface heuristic to be most accurate when it seeks variation within subfamilies of a protein family that have suffered only modest divergence. For example, the heuristic can be specified to consider only subfamilies of proteins that are less than 100 PAM units divergent.

Even this modified heuristic yields rather unsatisfactory surface predictions. For example, in alcohol dehydrogenases, the heuristic assigns 128 positions to the surface. Of these, however, only 96 (75%) are actually on the surface (with 50% side-chain exposure defining a surface residue, see above). As random predictions would be about 50% correct in this protein, the "naive" surface heuristic is not particularly satisfactory, at least in this case.

The failure of this heuristic to provide good surface predictions in alcohol dehydrogenase can be explained in evolutionary terms. The assumption that variation lies on the surface of a protein is grounded in the view that surface variation will not *alter the behavior of the protein sufficiently to* adversely influence the survival and reproduction of the host organism. In the language of molecular evolution, because there are fewer functional constraints on the divergence of surface residues, variation on the surface of a protein is more likely to be neutral (King & Jukes, 1969; Kimura, 1982), less likely to adversely influence the survival of the host organism, and therefore more likely to appear in a database.

Whatever the validity of this view, its application to surface prediction involves the logical converse, that all variation within a set of homologous protein sequences is neutral, a proposition that is clearly false. Superimposed on neutral variation is adaptive variation, which alters behavior within a protein family to make individual members better suited to perform different roles in different environments. In terms of biological function during divergent evolution, adaptive variation and neutral variation are opposites: adaptive variation alters behavior in a

Table 5

Surface prediction as a function of the number of variable subfamilies and the maximum PAM width of subfamilies

	Accuracy								Coverage							
MaxPW	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average
A. *One variable subfamily*																
010	100·0	94·1	91·6	100·0	100·0		100·0	97·6	6·6	9·4	25·8	22·3	13·0	0·0	4·9	11·7
020	100·0	89·3	91·6	100·0	87·1	100·0	100·0	95·4	7·0	24·7	25·8	34·0	49·2	19·3	28·3	26·9
040	91·4	86·5	84·8	96·0	81·0	100·0	100·0	91·4	15·0	34·1	32·9	52·1	68·1	37·0	43·2	40·3
060	91·4	86·0	84·8	96·3	79·6	100·0	100·0	91·2	15·0	43·5	32·9	56·3	73·9	45·1	44·4	44·4
080	93·2	86·5	85·5	90·9	79·4	100·0	97·6	90·4	19·2	45·2	34·7	63·8	78·2	45·1	51·8	48·3
100	93·2	85·8	85·5	90·9	79·4	100·0	97·7	90·4	19·2	50·0	34·7	63·8	78·2	45·1	53·0	49·1
120	82·5	85·8	85·5	90·9	79·4	91·3	97·7	87·6	66·2	50·0	34·7	63·8	78·2	67·7	54·3	59·3
140	82·5	85·8	78·0	90·9	79·4	91·3	97·7	86·5	66·2	50·0	56·4	63·8	78·2	67·7	54·3	62·4
160	82·5	85·8	78·0	90·9	74·3	91·3	97·7	74·4	66·2	50·0	56·4	63·8	88·4	67·7	54·3	63·8
180	82·5	85·8	78·0	90·9	74·3	91·3	81·0	83·4	66·2	50·0	56·4	63·8	88·4	67·7	79·0	67·4
200	82·5	69·1	78·0	84·6	74·3	91·3	81·0	80·1	66·2	77·6	56·4	81·9	88·4	67·7	79·0	73·9
B. *Two variable subfamilies*																
010	100·0	50·0	100·0	100·0	100·0	-	—	90·0	3·3	0·5	11·1	11·7	2·8	0·0	0·0	4·2
020	100·0	95·4	100·0	100·0	100·0	100·0	100·0	99·3	5·2	12·3	11·1	24·4	26·0	9·6	17·2	15·1
040	100·0	90·4	82·7	95·6	90·4	100·0	100·0	94·2	6·5	22·3	28·2	46·8	55·0	29·0	30·8	31·2
060	100·0	92·0	82·7	95·9	80·0	100·0	100·0	92·9	6·5	27·0	28·2	50·0	63·7	43·5	40·7	37·1
080	93·2	91·2	82·7	96·0	79·6	100·0	100·0	91·8	19·2	30·5	28·2	51·0	68·1	43·5	44·4	40·7
100	93·2	91·2	82·7	94·7	77·0	100·0	100·0	91·3	19·2	30·5	28·2	57·4	68·1	43·5	44·4	41·6
120	93·1	91·2	85·5	94·9	78·4	100·0	97·6	91·6	19·2	30·5	34·7	59·5	73·9	43·5	54·3	45·1
140	93·1	85·8	85·5	94·9	79·4	100·0	97·6	90·9	19·2	50·0	34·7	59·5	78·2	43·5	54·3	48·5
160	93·1	85·8	85·5	94·9	79·4	100·0	97·6	90·9	19·2	50·0	34·7	59·5	78·2	43·5	54·3	48·5
180	93·1	85·8	85·5	90·9	79·4	100·0	97·6	90·4	19·2	50·0	34·7	63·8	78·2	43·5	54·3	49·1
200	93·1	85·8	85·5	90·9	79·4	100·0	97·6	90·4	19·2	50·0	34·7	81·9	78·2	43·5	54·3	51·7
C. *Three variable subfamilies*																
010	100·0	100·0	100·0	100·0		—	—	100·0	0·5	0·5	4·1	7·4	0·0	0·0	0·0	1·8
020	100·0	85·7	100·0	100·0	100·0	100·0	100·0	98·0	0·9	3·5	4·1	14·8	14·4	6·4	8·6	7·5
040	100·0	95·6	94·1	100·0	90·0	100·0	100·0	97·1	0·9	12·9	18·8	27·6	43·4	9·6	16·0	18·5
060	100·0	96·2	94·1	100·0	85·1	100·0	100·0	82·2	0·9	15·2	18·8	30·8	57·9	9·6	29·6	23·3
080	100·0	92·3	94·1	100·0	80·3	100·0	100·0	95·2	2·3	21·1	18·8	36·2	65·2	9·6	29·6	26·1
100	100·0	92·3	94·1	100·0	78·3	100·0	100·0	95·0	2·3	21·1	18·8	40·4	68·1	9·6	29·6	27·1
120	100·0	92·3	94·1	100·0	78·3	100·0	100·0	95·0	2·3	21·1	18·8	40·4	68·1	9·6	29·6	27·1
140	100·0	92·3	94·1	100·0	78·3	100·0	100·0	95·0	2·3	21·1	18·8	40·4	68·1	9·6	29·6	27·1
160	100·0	92·3	94·1	100·0	78·3	100·0	100·0	95·0	2·3	21·1	18·8	40·4	68·1	9·6	29·6	27·1
180	100·0	92·3	94·1	100·0	78·3	100·0	100·0	95·0	2·3	21·1	18·8	40·4	68·1	9·6	29·6	27·1
200	100·0	92·3	94·1	100·0	78·3	100·0	100·0	95·0	2·3	21·1	18·8	40·4	68·1	9·6	29·6	27·1

Table 5 *continued*

MaxPW	Accuracy								Coverage							
	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average
D. *Four variable subfamilies*																
010	—	100·0	100·0	100·0	—	—	—	100·0	0·0	0·5	1·7	4·2	0·0	0·0	0·0	0·9
020	—	100·0	100·0	100·0	100·0	100·0	100·0	100·0	0·0	0·5	1·7	7·4	5·7	4·8	1·2	3·0
040	—	85·7	100·0	100·0	93·7	100·0	100·0	96·6	0·0	3·5	7·0	13·8	21·7	4·8	3·7	7·8
060	—	87·5	100·0	100·0	86·1	100·0	100·0	95·6	0·0	4·1	7·0	13·8	44·9	4·8	11·1	12·2
080	—	87·5	100·0	100·0	84·0	100·0	100·0	95·3	0·0	4·1	7·0	13·8	53·6	4·8	11·1	13·5
100	—	87·5	100·0	100·0	81·6	100·0	100·0	94·9	0·0	4·1	7·0	13·8	57·9	4·8	11·1	14·1
120	—	87·5	100·0	100·0	80·0	100·0	100·0	94·6	0·0	4·1	7·0	13·8	57·9	4·8	11·1	14·1
140	—	87·5	100·0	100·0	80·0	100·0	100·0	94·6	0·0	4·1	7·0	13·8	57·9	4·8	11·1	14·1
160	—	87·5	100·0	100·0	80·0	100·0	100·0	94·6	0·0	4·1	7·0	13·8	57·9	4·8	11·1	14·1
180	—	87·5	100·0	100·0	80·0	100·0	100·0	94·6	0·0	4·1	7·0	13·8	57·9	4·8	11·1	14·1
200	—	87·5	100·0	100·0	80·0	100·0	100·0	94·6	0·0	4·1	7·0	13·8	57·9	4·8	11·1	14·1

The maximum PAM width (MaxPW) for subfamilies is defined in the left column. Surface-indicating amino acids are KREND. The number of variable subfamilies required for a surface assignment to be made is indicated at the top of each section. Greater than 50% accessibility to a probe of 1·4 Å defines a surface residue. Accuracies for heuristics with zero coverage are designated by a dash. Averages are unweighted and have no exact interpretation. Average coverages include families where the heuristic identifies no surface residues; average accuracies exclude these. In some cases where zero averages are reported, this is due to an evolutionary tree that does not permit the designated number of subfamilies at the indicated MaxPW. Criterion A (Fig. 3) is used.

way that influences survival, while neutral variation does not. The structural implications are also opposite. Neutral variation must lie in the folded structure at positions that are the least important to behavior, primarily in regions of the surface that interact primarily with solvent (this does not mean, of course, that surface residues have no functional importance, or that certain surface residues do not perform critical functions). Adaptive variation, because it is intended to alter the behavior of the protein, can occur anywhere in a structure, including the active site.

Because neutral variation and adaptive variation look similar in an alignment of homologous protein sequences, efforts to extract conformational information from patterns of variation among homologous protein sequences should focus on strategies for distinguishing neutral from adaptive variation. This is especially true in protein families where the amount of adaptive variation is large (e.g. alcohol dehydrogenase; Jörnvall *et al.*, 1987; Benner, 1989).

(i) *A surface heuristic based on concurrent variation*

One approach for distinguishing neutral from adaptive variation relies on the notion of concurrent variation (Benner, 1989). Heuristics based on this approach assign a position to the surface if at least two subfamilies of the protein family, defined at a particular PAM distance, contain more than one type of residue at the position (the subfamilies are "variable"), and the variable subfamilies contain at least one amino acid chosen from the set Lys, Arg, Glu, Asn, and Asp (KREND). The notion is based on the hypothesis of independent variation (Benner, 1989; Benner, 1992*a*). This heuristic was used in the first *bona fide* predictions made in Zürich (Benner, 1989), where it was applied by hand. When applied automatically to the seven test protein families, accuracies ranging from 78 to 100% are obtained (Table 5, MaxPW ≥ 120). Coverages range from 19 to 82%.

These scores are dramatically influenced by the nature of the alignment, implying that averages over several protein families are not particularly informative; as in chemistry generally, individual cases must first be understood individually. For example, coverage increases and accuracy decreases with increasing numbers of proteins in the alignment. In these test cases, coverage in the phospholipase family (with 78 proteins) is remarkably high (73·9%) and accuracy remarkably low (78·5%), in contrast with the aspartate aminotransferase family (with only 17 proteins), where accuracy is remarkably high (93·1%), but coverage is low (only 19·2%).

These trends are not surprising, as an alignment containing a large number of proteins is likely to contain more subfamilies at any particular PAM width, these subfamilies are likely to contain more proteins, and therefore the subfamilies are more likely to have suffered variation. Further, more opportunities for variation (including compensated variation) are possible inside the folded structure with large protein families.

The balance of the evolutionary tree also influences the accuracy and coverage of the heuristics. For example, both the myoglobin and phospholipase families have 78 members. Yet the accuracy for the myoglobin family is higher (94·9%) and the coverage lower (59·5) than with phospholipase (MaxPW = 120). The myoglobin tree (Fig. 1(d)) is poorly balanced (compare the tree for phospholipase, Fig. 1(e)), and therefore has fewer subbranches containing more than one sequence diverging at relatively high PAM distances. This implies that coverage should generally be higher in protein families with more balanced trees than in families with less balanced trees with the same number of sequences.

(ii) *Generalizing the surface heuristic*

A large number of analogous surface heuristics can be generated by varying the heuristic parameters. First, the number of variable subfamilies can be increased. The evolutionary width of the subfamilies examined can be varied. The specified set of surface indicating hydrophilic amino acids can be altered. The distribution of these surface indicating residues across the alignment can be varied. In earlier work (in particular, Benner & Gerloff, 1991), a variety of these modified heuristics was applied by hand. Here, 2970 variants of the fundamental surface heuristic were tested systematically.

(ii) (a) *Changing the number of variable subfamilies*

"Concurrent variation" stipulates that an increase in the amount of variation is a more significant indicator of surface position when it is distributed across several subfamilies at a particular PAM distance, rather than being concentrated within a single subfamily. To test this notion systematically, a series of heuristics was constructed where surface assignments were made only if more than two, three, four, five, six, seven and eight subfamilies were variable at the designated position, and tested with each of the seven protein families.

The reliability of surface heuristics was observed to increase with increasing number of variable subfamilies regardless of the maximum PAM width of the subfamilies in which variation is sought (Table 5). For example, when 50% exposure is used to define a surface position, accuracies rise from about 88% when a surface assignment is made based on only a single variable subfamily to about 95% when a surface assignment is made based on three variable subfamilies, and approach 100% with five or more variable subfamilies (data not shown). Further, concurrent variation in two subfamilies at a MaxPW of 200 (Table 5) yields predictions that are 79 to 100% accurate, somewhat better than the 69 to 91% accuracy of surface assignments made if only one subfamily is variable. Great improvement in accuracy is seen in the alcohol dehydrogenase (ADH) family, which has sustained considerable divergence in function (especially substrate speci-

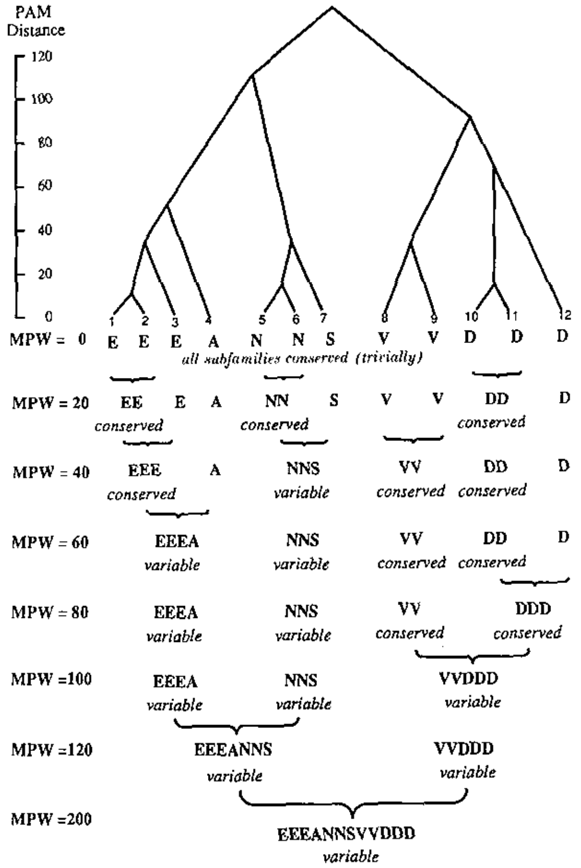

PAM Distance
120
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12
MPW = 0 E E E A N N S V V D D D
all subfamilies conserved (trivially)
MPW = 20 EE E A NN S V V DD D
conserved conserved conserved
MPW = 40 EEE A NNS VV DD D
conserved variable conserved conserved
MPW = 60 EEEA NNS VV DD D
variable variable conserved conserved
MPW = 80 EEEA NNS VV DDD
variable variable conserved conserved
MPW =100 EEEA NNS VVDDD
variable variable variable
MPW =120 EEEANNS VVDDD
variable variable
MPW =200 EEEANNSVVDDD
variable

Figure 2. A diagram constructing clusters of subfamilies of proteins at increasing PAM distance thresholds (MPW) for the purpose of assigning surface positions. Depicted is an idealized evolutionary tree of a protein family with 12 members. The amino acids present at a position in the multiple alignment are shown using the 1-letter code (A, alanine; D, aspartic acid; E, glutamic acid; N, asparagine; S, serine; V, valine). (a) At a MPW of 0, each protein is unconnected with any other protein. Therefore, no subfamily can be variable. At MPW = 20, proteins 1 and 2, proteins 5 and 6, and proteins 10 and 11 merge to form 3 subfamilies with more than 1 protein. All 3 subfamilies are conserved. No surface assignment is therefore made. (b) At MPW = 40, the connected component including proteins 1 and 2 adds a new member, protein 3. It remains conserved. The subfamily containing proteins 5 and 6 add protein 7, and becomes variable. Thus, at MPW = 40, the position would be assigned to the surface by an algorithm that specifies at least 1 variable subfamily, KREND, as surface-indicating amino acids, and a surface-indicating amino acid in a variable subfamily. It is not assigned to the surface if KRED are the surface-indicating amino acids. (c) At MPW = 60, the position is assigned to the surface even with KRED as the surface-indicating amino acids. The number of variable subfamilies $v = 1$ when subfamilies are counted as variable only if they contain a surface-indicating amino acid (distribution criterion A). The number of variable subfamilies $v = 2$ if subfamilies are counted as variable without regard to amino acid content (distribution criterion B). With KREND as the surface-indicating amino acids, $v = 2$. (d) At MPW = 80, all proteins belong to a subfamily with more than 1 member. (e) At MPW = 100, the position has 3 variable subfamilies, each with a surface-indicating amino acid (KREND), and is assigned to the surface by a surface algorithm with the following specifications ($v = 3$; MPW = 100, KREND).

ficity) and therefore (presumably) more adaptive variation (Jörnvall *et al.*, 1987). In contrast, in lactate dehydrogenase (LDH), where biological function has (again presumably) been constant throughout the divergent evolution represented by the alignment, the accuracy of the surface prediction with single subfamily or two subfamily variation is not markedly different.

(ii) (b) *Changing the maximum PAM width (MaxPW) of the subfamilies*

The structural significance of hydrophilic variation within a set of aligned homologous sequences depends on the overall evolutionary divergence among these sequences. With a multiple alignment containing y sequences defining an evolutionary tree containing $(y-2)$ 3-fold vertices, the sequences can be divided into $(y-1)$ clusters of subfamilies at different PAM distances (Fig. 2). Each set of subfamilies defines a cluster of subfamilies at the PAM distance of the highest node joining a pair of proteins within any subfamily. The yth grouping contains all of the sequences in the alignment.

At PAM 0, there are y subfamilies, each containing a single protein sequence. Proceeding up the tree to the PAM distance of the lowest node, a new cluster of $(y-1)$ subfamilies is defined, with one subfamily containing two protein sequences, the remaining containing a single sequence. This process can be followed until two subfamilies remain, each represented by a subtree, with each subtree joined by the 2-fold vertex at the top of the tree (Fig. 2).

A surface heuristic can be applied to each of these clusters of subfamilies. Obviously, in the first cluster of subfamilies with a maximum PAM width (MaxPW) of 0 PAM units, each subfamily contains a single sequence, no subfamily can display variability, and no position can be assigned to the surface by the heuristics described above. However, at higher MaxPW values, protein sequences come together to form subfamilies containing more than one sequence, variation within subfamilies is possible, and surface assignments begin to be made. The lower the MaxPW at which the constructed subfamilies display hydrophilic variation of any particular type, the stronger the surface assignment.

This procedure creates a progression of y surface heuristics applied to subfamilies with progressively higher maximum PAM widths. For convenience, these heuristics are grouped together in PAM "windows". Table 5 shows the accuracy and coverage of these heuristics as a function of the MaxPW of the variable subfamilies. As expected, accuracy generally decreases with increasing MaxPW. Nevertheless, accuracy remains quite variable among protein types. For example, with one variable subfamily in subfamilies defined by MaxPW = 200, the percentage of the surface assignments that are correct ranges from 69% (ADH) to 91% (PLC).

The most useful surface heuristic obtains the highest coverage at the greatest accuracy (that is, makes the most assignments with the fewest erroneous assignments). For example, with two variable subfamilies, upon going from MaxPW = 20 to MaxPW = 200, the average accuracy (unweighted) decreases from 99% to only 90%, at the same time as average coverage (unweighted) leaps from 15% to 52%. Thus, heuristics with high MaxPW values are generally more useful than heuristics with low MaxPW values. In practice, however, all heuristics can be used (see below).

Overall, heuristics that assign a third of the surface positions typically have accuracies of 85 to 90%. Those that assign 50% of the surface positions typically have accuracies of about 82%, with the lowest accuracy seen with LDH (80%) and the highest accuracy seen with MYO (96%).

(ii) (c) *Changing the definition of surface-indicating amino acid*

The surface heuristics discussed so far prescribe that each variable subfamily, to be counted as variable, must contain a surface-indicating amino acid. In the original work (Benner, 1989), five amino acids were defined as surface-indicating based on a combination of intuition and empirical information: Lys, Arg, Glu, Asn and Asp (KREND). Here, the definition of surface-indicating residue was systematically varied to yield another set of surface heuristics, and each evaluated using the seven test protein families. A summary of the data is found in Table 6. The data show several trends. First, the more amino acids included in the definition of surface-indicating, the greater the coverage and the lower the accuracy. However, the change in accuracy is not equal in all protein families. For example, at MaxPW = 100 and with two variable subfamilies, accuracy suffers both with the alcohol dehydrogenase and myoglobin families when the set of amino acids defined as surface-indicating is expanded. Hardly any accuracy is lost, in contrast, with the phospholipase and aspartate aminotransferase families.

(ii) (d) *Changing the prescribed distribution of surface-indicating amino acids*

The distribution of surface-indicating residues among subfamilies is also a parameter of surface heuristics. Scores for three different surface heuristics prescribing three different types of distribution of surface-indicating residues, illustrated in Figure 3, are compiled in Table 7.

In the first distribution (criterion A), a subfamily contributes to the tally of variable subfamilies if (and only if) it contains both more than one type of residue and at least one surface-indicating amino acid. The position is assigned to the surface if the position contains the prescribed number of variable subfamilies (or more).

In the second distribution (criterion B), a subfamily contributes to the tally of variable subfamilies if it contains more than one residue type, regardless of whether it contains a surface-indicating amino acid. The position may be assigned to the surface if it contains the prescribed number of

Table 6

Accuracy and coverage of surface heuristics depending on the definition of surface-indicating amino acid

Amino acid	Accuracy								Coverage							
	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average	AAT	ADH	LDH	MYO	PLA	PLC	SOD	Average
KRED	91·8	92·0	82·6	96·1	77·0	100·0	100·0	91·4	15·9	27·0	25·2	53·1	68·1	38·7	39·5	38·2
KREND	93·1	91·2	82·7	94·7	77·0	100·0	100·0	91·2	19·2	30·5	28·2	57·4	68·1	43·5	44·4	41·6
KRENDQ	92·1	90·6	82·2	90·4	77·4	100·0	100·0	90·4	22·0	34·1	30·0	60·6	69·5	45·1	48·1	44·2
KRENDH	93·4	88·7	77·7	91·8	77·7	100·0	100·0	89·9	20·1	32·3	28·8	59·5	71·0	45·1	44·4	43·0
KRENDS	92·1	90·6	82·2	90·4	77·4	100·0	100·0	90·4	22·0	34·1	30·0	60·6	69·5	45·1	48·1	44·2
KRENDHQ	92·1	88·4	78·7	90·7	78·1	100·0	100·0	89·7	22·0	35·8	30·5	62·7	72·4	46·7	48·1	45·5
KRENDSH	91·8	81·9	74·2	87·8	79·4	100·0	97·6	87·5	26·2	40·0	30·5	69·1	78·2	54·8	51·8	50·1
KRENDSQ	90·7	82·9	78·2	86·4	79·4	100·0	97·7	87·9	27·6	40·0	31·7	68·0	78·2	54·8	54·3	50·7
KRENDQST	89·3	79·5	77·3	85·0	80·0	100·0	96·0	86·7	27·6	41·1	34·1	72·3	81·1	56·4	59·2	53·1
KRENDSHQ	90·7	81·6	75·3	86·8	79·4	100·0	97·7	87·4	27·6	41·7	32·3	70·2	78·2	56·4	54·3	51·5

Greater than 50% accessibility to a probe of 1·4 Å defines a surface residue. Data collected for subfamilies defined by a maximum PAM width of 100. Two variable subfamilies indicate a surface position. Criterion A (see Fig. 3) was used to define the distribution of surface-indicating amino acids. Averages are unweighted and have no exact interpretation.

Alignment	Criterion A	Criterion B	Criterion C
EEES AAAK QQSS	2 variable surface	3 variable surface	3 variable surface
EEES AAAK QQSK	3 variable surface	3 variable surface	3 variable surface
EEES AAAK QQQQ	2 variable surface	2 variable surface	2 variable surface
AASS AAQQ KKKK	0 variable *no assignment*	2 variable *no assignment*	2 variable *surface*

Figure 3. Illustration of 3 distribution criteria for surface-indicating amino acids in a surface algorithm. Depicted are hypothetical sequences at a single position in a protein family with 12 members. At the designated MPW, the 12 proteins form 3 subfamilies, each with 4 members. Surface-indicating amino acids are KREND. Criterion A: surface-indicating amino acid must be within subfamily for it to count as variable. Criterion B: subfamily is variable if it contains more than one amino acid, regardless of the amino acids it contains. The position is assigned to the surface only if at least one of the variable subfamilies contains a surface-indicating amino acid. Criterion C: the position is assigned to the surface provided the prescribed number of variable subfamilies are observed, and a surface-indicating amino acid is found in any subfamily, variable or conserved.

Table 7

Criteria for the distribution of surface-indicating residues in surface assignment heuristics

	Accuracy			Coverage		
A. *Aspartate aminotransferase (AAT)*						
	A	B	C	A	B	C
v1	93·1	91·4	91·4	19·2	20·1	20·1
v2	93·1	92·8	92·8	19·2	30·5	30·5
v3	100·0	92·8	92·8	2·3	6·1	6·1
v4	0·0	0·0	0·0	0·0	0·0	0·0
v5	0·0	0·0	0·0	0·0	0·0	0·0
v6	0·0	0·0	0·0	0·0	0·0	0·0
B. *Alcohol dehydrogenase (ADH)*						
v1	85·8	81·0	81·0	50·0	47·6	47·6
v2	91·2	83·1	83·1	30·5	43·5	43·5
v3	92·3	83·3	83·3	21·1	38·2	38·2
v4	87·5	82·6	82·6	4·1	11·1	11·1
v5	100·0	100·0	100·0	0·5	1·1	0·5
v6	100·0	100·0	100·0	0·5	0·5	0·5
C. *Lactate dehydrogenase (LDH)*						
v1	85·5	86·5	86·5	34·7	34·1	34·1
v2	82·7	83·8	83·8	28·2	33·5	33·5
v3	94·1	86·0	86·0	18·8	25·2	25·2
v4	100·0	86·3	86·3	7·0	11·1	11·1
v5	100·0	75·0	75·0	1·1	1·7	1·7
v6	0·0	0·0	0·0	0·0	0·0	0·0
D. *Myoglobin (MYO)*						
v1	90·9	90·7	90·7	63·8	52·1	52·1
v2	94·7	89·7	89·7	57·4	64·8	64·8
v3	100·0	89·6	89·6	40·4	55·3	55·3
v4	100·0	92·0	92·0	13·8	24·4	24·4
v5	100·0	100·0	100·0	5·3	11·7	11·7
v6	100·0	100·0	100·0	2·1	7·4	7·4
E. *Phospholipase (PLA)*						
v1	79·4	80·3	80·3	78·2	59·4	59·4
v2	77·0	73·3	73·3	68·1	63·7	63·7
v3	78·3	76·9	76·9	68·1	72·4	72·4
v4	81·6	78·4	78·4	57·9	73·9	73·9
v5	84·6	78·9	78·9	47·8	65·2	65·2
v6	85·1	81·2	81·2	33·3	56·5	56·5

Table 7 *continued*

	Accuracy			Coverage		
F. *Plastocyanin (PLC)*						
v1	100·0	95·4	95·4	45·1	33·8	33·8
v2	100·0	97·0	97·0	43·5	53·2	53·2
v3	100·0	95·0	95·0	9·6	30·6	30·6
v4	100·0	100·0	100·0	4·8	16·1	16·1
v5	100·0	100·0	100·0	3·2	9·6	9·6
v6	100·0	100·0	100·0	1·6	3·2	3·2
G. *Superoxide dismutase (SOD)*						
v1	97·7	96·0	96·0	53·0	60·4	60·4
v2	100·0	97·7	97·7	44·4	54·3	54·3
v3	100·0	100·0	100·0	29·6	43·2	43·2
v4	100·0	100·0	100·0	11·1	27·1	27·1
v5	100·0	100·0	100·0	1·2	12·3	12·3
v6	100·0	100·0	100·0	1·2	1·2	1·2

The maximum PAM width for subfamilies is 100. Surface-indicating amino acids are KREND. The number of variable subfamilies required for a surface assignment is indicated in the left-hand column. Greater than 50% accessibility to a probe of 1·4 Å defines a surface residue. Criterion A: a subfamily is considered to contribute to hydrophilic variation at a position if and only if it contains both more than 1 type of amino acid and at least 1 surface-indicating amino acid. The position is assigned to the surface if it contains the indicated number of variable subfamilies (or more). Criterion B: a subfamily is again counted as variable if it contains more than 1 amino acid, regardless of whether it contains a surface-indicating amino acid. The position is assigned to the surface, however, only if at least 1 of the variable subfamilies contains a surface-indicating amino acid. Criterion C: a subfamily is counted as variable if it contains more than 1 amino acid, regardless of whether it contains a surface-indicating amino acid. Again, the position is assigned to the surface if it contains the indicated number of variable subfamilies (or more). The position is assigned to the surface, however, if at least 1 protein at that position contains a surface-indicating amino acid, regardless of whether or not that protein is part of a variable subfamily.

variable subfamilies (or more), and only if at least one of the sequences in one of the variable subfamilies contains a surface-indicating amino acid.

In the third distribution (criterion C), a subfamily contributes to the tally of variable subfamilies if it contains more than one residue type, regardless of whether it contains a surface-indicating amino acid. Again, the position must contain the prescribed number of variable subfamilies (or more) to be assigned to the surface. The position is assigned to the surface, however, if at least one protein at that position contains a surface-indicating amino acid, regardless of whether or not that amino acid is in a protein in one of the variable subfamilies.

Distribution criterion A is more stringent than criterion B, which is more stringent than criterion C. Criterion C assigns the most positions to the surface, but is expected to have the lowest accuracy. Criterion A is expected to assign the fewest positions to the surface with the greatest accuracy. These trends were observed in all proteins at all PAM distances and with all definitions of surface-indicating amino acids (data not shown).

(b) *Interior heuristics*

(i) *Theory of interior assignments*

In water-soluble proteins, hydrophobic side-chains are found preferentially inside the folded structure (Schulz & Schirmer, 1979), and the literature contains many proposals for identifying interior amino acids from their hydrophobicity (Schiffer & Edmunson, 1967; Lim, 1974*a*,*b*; Kyte & Doolittle, 1982; Eisenberg *et al.*, 1982; Kaiser & Kezdy, 1984). The preference is, however, far from absolute (Lee & Richards, 1971). Natural proteins have many hydrophobic side-chains on the surface, where they may form contacts with other proteins, modulate the solubility of the protein, or simply violate folding rules to obtain proteins with a level of conformational instability desired by natural selection.

To the extent that the last explanation is true, interior assignments (and secondary structure predictions derived from them) should be improvable by averaging heuristics over a set of aligned homologous protein sequences. Further, residues whose side-chains lie inside are presumably subject to greater functional constraints on divergence than residues on the surface, as changes within a packed interior of a protein are less likely to be accepted by natural selection. Thus, combining generalization (1) (hydrophobic residues lie inside) and generalization (3) (interior residues are more highly conserved) suggested interior heuristics that were used to build *bona fide* predictions for several protein families (Benner, 1989; Benner & Gerloff, 1991; Benner, 1992*b*; Benner *et al.*, 1992; 1993*a*; Gerloff *et al.*, 1993*a*,*b*), again applied by hand. Here, we automatically and systematically generate a set of interior heuristics and evaluate them for the set of test protein families described above.

(ii) *Two simple interior heuristics*

Two simple interior heuristics can be proposed. The first requires that the same interior-indicating residue, a hydrophobic amino acid such as Phe, Ala, Met, Ile, Leu, Tyr, Val or Trp (FAMILYVW, see below), be at the designated position in every protein in an alignment. The second requires simply that some interior-indicating residue be present at the designated position in all proteins in the alignment; different interior-indicating residues can be contributed by different proteins.

The accuracies and coverages of these two heuristics applied to the test protein families are shown in Table 8 and Table 9. The first heuristic, which requires that the same hydrophobic residue be found in all proteins in the alignment, is clearly more stringent than the second, which requires only that some hydrophobic residue be found in all proteins in the alignment. For the first heuristic, coverages range widely, from 0·5 to 20% (when <50% exposure defines an interior residue). The extent of coverage varies roughly inversely with the PAM width of the alignment; the higher the overall divergence in the protein family, the less likely that any specific residue is conserved over the entire alignment. Accuracies also vary widely (from 40% to 100%) and are surprisingly low, especially when compared with those routinely obtained with surface heuristics.

Table 8

Identifying interior positions. All proteins have the same interior-indicating residue (FAMILYVW)

	<50% Exposure		<40% Exposure		PAM
Protein	*Accuracy*	*Coverage*	*Accuracy*	*Coverage*	*width*
AAT	71·0	12·0	74·2	10·6	104
ADH	100·0	0·5	100·0	0·4	190
LDH	53·3	5·1	66·7	5·6	135
MYO	66·7	3·4	100·0	4·2	190
PLA	100·0	5·9	100·0	5·3	160
PLC	77·8	20·0	100·0	19·1	102
SOD	40·0	2·9	40·0	2·4	163
Average	75·9	7·0	83·0	6·8	

Accuracies and coverages of a heuristic that assigns a position to the interior if all sequences have the same amino acid, and where that amino acid is one of the following: FAMILYVW. Examined at 2 different definitions of interior residue, less than 50% side-chain exposure and less than 40% side-chain exposure. Averages are unweighted and have no exact interpretation.

For the second heuristic, coverages are between 15 and 47%, with a less clear relationship between coverage and the PAM width of the alignment. Coverages do not appear to depend strongly on the PAM width of the alignment overall. Remarkably, the accuracy of this heuristic is not necessarily less. This is the first of a series of observations that suggests that hydrophobicity coupled with *variation* leads to a stronger interior heuristic than hydrophobicity coupled with conservation (see below).

(iii) *Modifying interior heuristics*

The parameters of interior heuristics were then systematically altered and automatically evaluated in the seven test protein families. Parameters varied were: the nature of the "interior indicating" hydrophobic amino acid, the distribution and extent of conservation, and the maximum PAM width used to define the subfamilies within which conservation is observed.

Table 9

Identifying interior positions. All proteins have an interior-indicating residue (FAMILYVW)

	<50% Exposure		<40% Exposure		PAM
Protein	Accuracy	Coverage	Accuracy	Coverage	width
AAT	75·0	31·1	85·5	30·1	104
ADH	85·7	23·5	87·5	21·3	190
LDH	74·6	31·8	79·1	29·9	135
MYO	87·1	46·6	93·5	40·3	190
PLA	72·7	15·7	72·7	14·0	160
PLC	76·9	28·6	92·3	25·5	102
SOD	100·0	22·1	100·0	18·3	163
Average	81·7	28·5	87·2	25·6	

Accuracies and coverages of a heuristic that assigns a position to the interior if all sequences have an amino acid chosen from the set FAMILYVW, but no specific amino acid is conserved across the entire alignment. Examined at 2 different definitions of interior residue, less than 50% side-chain exposure and less than 40% side-chain exposure. Averages are unweighted and have no exact interpretation.

Table 10

Identifying interior positions. All proteins have the same interior-indicating residue (FAMILYVWPG)

	<50%		<40%		PAM
Protein	*Accuracy*	*Coverage*	*Accuracy*	*Coverage*	*width*
AAT	72·5	20·2	80·3	19·0	104
ADH	66·7	4·9	66·7	0·4	190
LDH	54·8	10·8	61·3	10·7	135
MYO	75·0	5·2	100·0	5·6	190
PLA	100·0	9·8	100·0	8·8	160
PLC	50·0	8·6	70·0	29·8	102
SOD	57·1	11·8	64·3	11·0	163
Average	68·0	10·2	77·5	12·2	

Accuracies and coverages of a heuristic that assigns a position to the interior if all sequences have the same amino acid, and where that amino acid is one of the following: FAMILYVWPG. Examined at 2 different definitions of interior residue, less than 50% side-chain exposure and less than 40% side-chain exposure. Averages are unweighted and have no exact interpretation.

(iii) (a) *Changing the definition of interior-indicating amino acid*

In the simple interior heuristics outlined above, the residues FAMILYVW are considered interior-indicating. A new set of heuristics was evaluated with this set expanded to include Pro and Gly. The scores are collected in Table 10 (where a single specific interior-indicating residue is conserved across the entire alignment) and Table 11 (where any one of the interior-indicating residues is found at the designated position in all proteins, but with no specific residue conserved across the entire alignment). Coverages were uniformly higher, as expected given the larger number of interior-indicating amino acids. Accuracies are modestly to substantially less.

It should be noted that a Pro or a Gly conserved across an entire alignment also is a parsing element (Benner, 1989; Benner & Gerloff, 1991). It serves to divide segments of the alignment into manageable units that are separately assigned secondary struc-

Table 11

Identifying interior positions. All proteins have an interior-indicating residue (FAMILYVWPG)

	<50%		<40%		PAM
Protein	Accuracy	Coverage	Accuracy	Coverage	width
AAT	70·1	33·3	80·5	32·4	104
ADH	82·7	32·8	85·2	30·0	190
LDH	70·5	35·0	75·6	33·3	135
MYO	84·8	48·3	93·9	43·1	190
PLA	76·9	19·6	76·9	17·5	160
PLC	75·0	34·3	87·5	29·8	102
SOD	95·7	32·4	100·0	28·0	163
Average	79·4	33·7	85·7	30·6	

Accuracies and coverages of a heuristic that assigns a position to the interior if all sequences have amino acids chosen from the set FAMILYVWPG, but no specific amino acid is conserved across the entire alignment. Examined at 2 different definitions of interior residue, less than 50% side-chain exposure and less than 40% side-chain exposure. Averages are unweighted and have no exact interpretation.

ture. Thus, expanding the definition of "hydrophobic" to include P and G causes interior assignments to be made to positions that will also be identified as parses in a separate phase of a structure prediction effort.

A still broader interior heuristic was examined, where all residues in all sequences at the designated position are chosen from the set CHQSTFAMILYVWPG, but where no single residue type is conserved across the entire alignment. The heuristic is equivalent to one where an interior assignment is made if the position lacks one of the five hydrophilic amino acids KREND. Data for this heuristic are collected in Table 12A. Accuracies are still less and coverages still higher. Such positions are discussed further below.

Positions where a specific amino acid chosen from the group CHQST is absolutely conserved across an entire alignment are candidates for an active site assignment (Benner, 1989; Benner & Gerloff, 1991). As discussed in these earlier publications, the strength of an active site assignment depends in part on the context in which it is found. Therefore, such positions are not normally assigned either to the surface or to the interior of the folded structure.

(iii) (b) *Changing the pattern of conservation*

In positions where hydrophobicity is conserved, but where no single residue is conserved across the entire alignment, the pattern of conservation provides additional structural information. As with surface heuristics, the distribution of conservation is defined with respect to subfamilies constructed with different PAM widths. There are two extremes (Fig. 4). In the first, every subfamily at a position may contain only one hydrophobic amino acid conserved within the subfamily (a "hydrophobic split" position). At the other, each subfamily may be variable, containing more than one hydrophobic amino acid (a "hydrophobic variable" position). Heuristics that assign interior positions by identifying hydrophobic splits and hydrophobic variable positions were systematically evaluated using the test set of proteins.

Alignment	Designation
VVVV AAAA FFFF	Hydrophobic split
VVVV AAAA CCCC	Non-hydrophilic split
CCCC HHHH SSSS	Neutral split
VVVV AAAA RRRR	Amphiphilic split
KKKK HHHH TTTT	Non-hydrophobic split
KKKK DDDD NNNN	Hydrophilic split
PPPP GGGG GGGG	Parsing split
VVAA FVAA FFFF	Hydrophobic Variable, 2 variable subgroups
VVAA FVAA FFFY	Hydrophobic Variable, 3 variable subgroups
VVTT SSAA FFFF	Non-hydrophilic Variable, 2 variable subgroups
VVTT SSAA FFFV	Non-hydrophilic Variable, 3 variable subgroups

Figure 4. Illustration of splits and hydrophobic variable positions. Illustration of 3 distribution criteria for surface-indicating amino acids in a surface algorithm. Depicted are hypothetical sequences at a single position in a protein family with 12 members. At the designated MPW, the 12 proteins form 3 subfamilies, each with 4 members. Surface-indicating amino acids are KREND. Interior-indicating amino acids are FAMILYVW. Neutral amino acids are CHQST.

(iii) (b) (i) *Hydrophobic split*

Heuristics identifying hydrophobic splits are the reciprocal of heuristics used to identify surface positions in an alignment. These focus on conservation (rather than variation) of hydrophobic (rather than hydrophilic) amino acids in subfamilies defined at

Table 12
Interior assignments with all positions non-hydrophilic

	<50%		<60%		PAM
Protein	Accuracy	Coverage	Accuracy	Coverage	Width
A. *Non-hydrophilic = CHQSTPG*					
AAT	66·7	35·0	74·0	32·9	104
ADH	75·0	44·1	83·3	43·5	190
LDH	64·4	36·9	68·9	35·0	135
MYO	60·9	24·1	69·6	22·2	190
PLA	77·3	33·3	81·8	31·6	160
PLC	45·8	31·4	58·3	29·8	102
SOD	72·7	47·1	81·8	43·9	163
Average	66·1	36·0	74·0	34·1	
B. *Non-hydrophilic = CHQST*					
AAT	69·2	24·6	73·8	22·2	104
ADH	76·5	30·4	87·7	30·9	190
LDH	69·8	28·0	74·6	26·5	135
MYO	60·0	20·7	65·0	18·1	190
PLA	72·2	25·5	77·8	24·6	160
PLC	60·0	17·1	70·0	14·9	102
SOD	70·4	27·9	77·8	25·6	163
Average	68·3	24·9	75·2	23·3	

Accuracies and coverages of a heuristic that assigns a position to the interior if all sequences have a non-hydrophilic amino acid, but where no specific amino acid is conserved across the entire alignment. Examined at 2 different definitions of interior residue, less than 50% side-chain exposure and less than 60% side-chain exposure. Averages are unweighted and have no exact interpretation.

Table 13

Hydrophobic splits. Interior prediction as a function of maximum PAM width

Max	AAT		ADH		LDH		MYO		PLA		PLC		SOD	
PW	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
A. *Interior-indicating amino acids are FAMILYVW*														
000+	36·3	2·2	73·9	8·3	40·0	2·5	50·0	1·7	100·0	3·9	66·6	5·7	100·0	8·8
010+	30·0	1·6	62·5	4·9	28·5	1·3	100·0	1·7	100·0	3·9	50·0	2·8	100·0	7·3
020+	42·8	1·6	58·3	3·4	28·5	1·3			100·0	2·0	50·0	2·8	100·0	5·9
040+	50·0	1·1	50·0	1·5	28·5	1·3					50·0	2·8	100·0	2·9
060+	50·0	1·1	50·0	1·5	33·3	1·3					0·0	0·0	100·0	1·5
080+	0·0	0·0	75·0	1·5	50·0	0·6							100·0	1·5
100+			100·0	1·0	50·0	0·6							0·0	0·0
120+			100·0	1·0										
B. *Interior-indicating amino acids are FAMILYVWPG*														
000+	73·5	42·6	85·9	24·0	69·2	28·7	85·2	50·0	78·5	21·6	77·2	48·6	85·0	25·0
010+	75·2	35·0	88·8	15·7	60·0	11·5	82·1	39·6	78·5	21·6	75·0	42·8	84·2	23·5
020+	73·6	30·6	86·6	12·7	60·0	11·5	80·9	29·3	76·9	19·6	76·4	37·1	80·0	17·6
040+	67·2	21·3	86·6	6·4	60·0	9·5	76·4	22·4	80·0	15·7	80·0	34·3	62·5	7·3
060+	67·2	21·3	80·0	3·9	59·0	8·3	71·4	17·2	87·5	13·7	80·0	22·8	50·0	4·4
080+	69·7	16·4	71·4	2·4	52·6	6·4	57·1	6·9	100·0	5·9	77·7	20·0	50·0	4·4
100+	70·9	12·0	33·3	0·5	52·6	6·4	57·1	6·9	100·0	5·9	77·7	20·0	40·0	2·9
120+			33·3	0·5	53·3	5·1	75·0	5·2	100·0	5·9			40·0	2·9
140+			100·0	0·5			75·0	5·2	100·0	5·9			40·0	2·9
160+			100·0	0·5			75·0	5·2					40·0	2·9
180+			100·0	0·5			66·6	3·4						

A hydrophobic split is a position where a particular amino acid is conserved within subfamilies, but not necessarily between subfamilies, defined at a particular PAM width. To be recorded at a particular row, the split must be observed *in subfamilies having the PAM width indicated in the left column* or higher. Less than 50% accessibility to a probe of 1·4 Å defines an interior residue.

increasing maximum PAM widths (MaxPW). The reliability of the interior assignment increases (as opposed to decreases) at increasing MaxPW. The process of searching for the highest MaxPW where a split is observed (as opposed to the lowest MaxPW where multiple variable subfamilies are observed) is illustrated in Figure 5. At MaxPW = 0, each subfamily contains only a single sequence, and no variation is possible. Thus, all positions are trivially designated as splits in this cluster. With increasing maximum PAM widths, however, subfamilies grow to include more proteins. At some point, one of the subfamilies becomes variable, and the position is no longer designated as a split. In Figure 5, the split is "lost" between PAM 40 and PAM 60 due to the addition of sequence 4 into the subfamily containing sequences 1, 2, and 3.

Data for the accuracy and coverage of splits were collected at various PAM windows. These are collected in Table 13, where both FAMILYVW and FAMILYVWPG are used as the interior-indicating amino acids. A hydrophobic split at a MaxPW equal to the PAM width of the alignment is, of course, simply an absolutely conserved hydrophobic residue, as the "subfamily" defined by a MaxPW equal to the PAM width of the entire alignment includes the entire alignment. Thus, the scores are the same as the scores in Tables 8 and 10. At lower MaxPW values, the heuristic makes more assignments.

Accuracies in general are higher at increasing MaxPW values where all subfamilies remain non-variable. Remarkably, however, the accuracy of this interior heuristic is not a steadily increasing function of MaxPW distance in all protein families (Table 13). These results are among the most surprising in this work.

Splits were then examined where the conserved amino acids were not constrained to a specific set of amino acids, according to the systematic outlined in Table 14. The accuracies and coverages of heuristics based on an analysis of splits are collected in Table 15. Neutral splits involving only the amino acids CHQST (and no other amino acids) are generally good indicators of interior position; in many proteins, the accuracy is as great as that predicted by hydrophobic splits. Splits involving both CHQST and PG are infrequent, and no structural conclusion can be drawn from the small number of examples available.

Amphiphilic splits, where at least one subfamily of proteins has a conserved FAMILYVW and at least one subfamily has a conserved KREND, generally contain amino acid whose side-chains lie on the surface of a protein in the representative crystal structure in the seven test families of proteins. The matching of a conserved hydrophobic residue against a conserved hydrophilic residue in an alignment is also noteworthy on other grounds. At large PAM distances, an amphiphilic split often indicates that the conformations of proteins in the two subfamilies are different. Such positions are therefore often useful in confirming assignments of secondary structure, as illustrated at key points in the prediction of the structure of protein kinase (Benner & Gerloff, 1991).

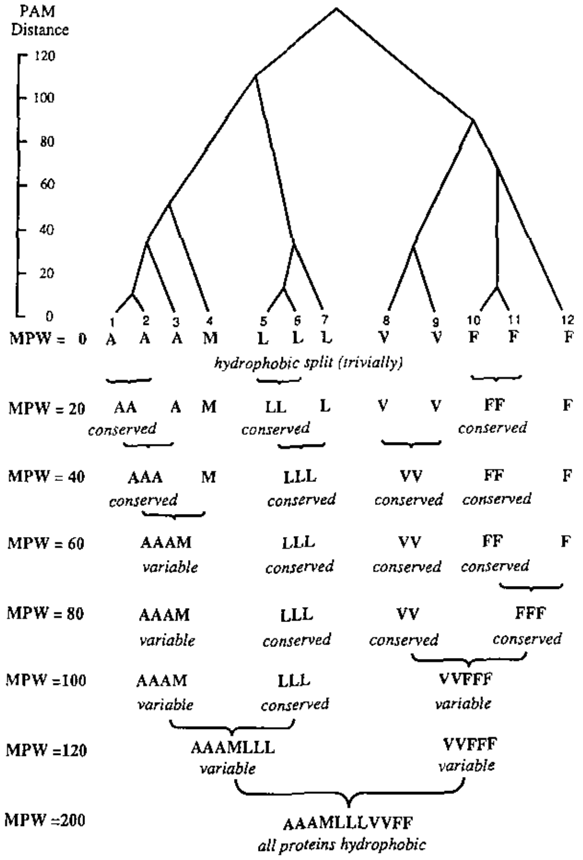

PAM
Distance
120
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12
MPW = 0 A A A M L L L V V F F F
hydrophobic split (trivially)
MPW = 20 AA A M LL L V V FF F
conserved conserved conserved
MPW = 40 AAA M LLL VV FF F
conserved conserved conserved conserved
MPW = 60 AAAM LLL VV FF F
variable conserved conserved conserved
MPW = 80 AAAM LLL VV FFF
variable conserved conserved conserved
MPW =100 AAAM LLL VVFFF
variable conserved variable
MPW =120 AAAMLLL VVFFF
variable variable
MPW =200 AAAMLLLVVFF
all proteins hydrophobic

Figure 5. A diagram constructing clusters of subfamilies of proteins at increasing PAM distance thresholds (MPW) for the purpose of assigning interior positions. Depicted is an idealized evolutionary tree of a protein family with 12 members. The amino acids present at a position in the multiple alignment are shown using the 1-letter code (A, alanine; F, phenylalanine; L, leucine; M, methionine; V, valine). (a) At a MPW of 0, each protein is unconnected with any other protein. Therefore, no subfamily can be variable, and the position is (trivially) a hydrophobic split. (b) At MPW = 20, proteins 1 and 2, proteins 5 and 6, and proteins 10 and 11 merge to form 3 subfamilies with more than 1 protein. All 3 subfamilies are conserved. The position is designated a hydrophobic split at MPW = 20. (c) At MPW = 40, the connected component including proteins 1 and 2 adds a new member, protein 3. It remains conserved. The subfamily containing proteins 5 and 6 adds protein 7, and remains conserved. The position remains a hydrophobic split, but at a higher MPW, 40. (d) At MPW = 60, the connected component that includes proteins 1, 2 and 3 adds a new member, protein 4, and becomes variable. The position can no longer be designated a split. It is now designated a hydrophobic variable position, with $v = 1$. (e) At MPW = 80, the position remains a hydrophobic variable position, with $v = 1$. (f) At MPW = 100, the position has 2 variable subfamilies. It is now designated a hydrophobic variable position, with $v = 2$. (g) At MPW >150, all of the proteins are included in a single subfamily. The position overall is designated "all positions interior-indicating".

Table 14
Logic table defining types of splits based on distribution of residue types

Description	FAMILYVW	CHQST	KREND	PG	Example
Hydrophobic	+	−	−	−	FFF VVV LLL MMM
Hydrophobic	+	−	−	+	FFF VVV GGG MMM
Non-hydrophilic	+	+	−	−	FFF CCC TTT VVV
Non-hydrophilic	+	+	−	+	FFF CCC GGG YYY
Neutral	−	+	−	−	CCC SSS QQQ HHH
Neutral	−	+	−	+	CCC SSS GGG PPP
Amphiphilic	+	−	+	−	FFF VVV KKK EEE
Amphiphilic	+	−	+	+	FFF KKK GGG EEE
Amphiphilic	+	+	+	−	FFF HHH KKK QQQ
Amphiphilic	+	+	+	+	FFF KKK CCC PPP
Non-hydrophobic	−	+	+	−	HHH SSS KKK EEE
Non-hydrophobic	−	+	+	+	HHH KKK PPP EEE
Hydrophilic	−	−	+	−	KKK DDD RRR EEE
Hydrophilic	−	−	+	+	KKK DDD GGG PPP
Parsing	−	−	−	+	PPP PPP GGG PPP

A + indicates that at least 1 subfamily contains at least 1 of the set of amino acids designated at the top of the column. A − indicates that no subfamily contains 1 of the set of amino acids designated at the top of the column. The example shows the amino acid residues present in a hypothetical protein family with 4 subfamilies each containing 3 sequences.

Non-hydrophobic splits (positions where the conserved amino acid is any other than FAMILYVW) and hydrophilic splits (containing only KREND and PG) are generally good indicators of surface positions. At very high MaxPW values, such splits suggest a functional constraint on divergence that is greater than expected for a normal surface position. Thus, non-hydrophobic splits at very high MaxPW values can indicate a position near the active site, or in the interior. This can be seen in Table 15, where hydrophilic splits at very large PAM distances do indicate interior position with reasonable accuracy. Again, absolutely conserved functionalized amino acids (such as KREND) often indicate an active-site position (Zvelebil & Sternberg, 1988), depending on the context and the PAM width of the alignment being inspected (Benner, 1989; Benner & Gerloff, 1991).

Another surprising conclusion to be derived from this analysis is that parsing splits involving only P and G are reasonably good indicators of an interior position. This outcome may be due in part to the use of backbone atoms in the calculation of surface accessibility only in the case of Gly. Nevertheless, such splits are used as parsing elements in the Zurich method (Benner, 1989; Benner & Gerloff, 1991). That parsing splits are often inside is surprising because parses normally involve turns or coils in the protein structure, elements that generally lie on the surface of the folded structure.

(iii) (b) (ii) *Hydrophobic variable*

The observation that hydrophobic splits at large PAM distances are not necessarily good indicators of interior positions led us to examine other filters to improve the accuracy and coverage of interior heuristics based on patterns of conservation and variation of hydrophobic groups. Subfamilies defined at particular MaxPW can be variable and contain only hydrophobic amino acids. Such positions are termed "hydrophobic variable". As with surface heuristics, heuristics that detect hydrophobic variability are characterized by the number of variable subfamilies and the maximum PAM width of the subfamily where this number of variable subfamilies is observed.

Scores from a set of heuristics seeking hydrophobic variability are collected in Tables 16 and 17. These heuristics make remarkably accurate interior assignments, most notably in alignments with large PAM widths and many proteins. Coverages of over 30% and accuracies over 90% are not uncommon.

As noted above, it appears that hydrophobicity and certain types of variation are together more reliable indicators of interior positions than hydrophobicity and conservation. This implies that if variation can be tolerated because a position lies on the surface, there is a high probability that a hydrophilic residue both can be and will be incorporated at this position, even after a small amount of divergent evolution. Conversely, if variation is observed and such a hydrophilic residue is not incorporated, this is a strong indication that the position is interior. However, an absolutely conserved hydrophobic residue may indicate functional constraints on divergence other than simply an interior position. Regardless of the underlying explanation, the hydrophobic variable heuristic is among the most useful to assign interior positions (Benner & Gerloff, 1991).

(iii) (b) (iii) *Non-hydrophilic variable*

A final set of heuristics was evaluated where variable positions contain the amino acids CHQST and FAMILYVW(PG), but not KREND. These are not assigned to the surface by the most useful surface heuristics, and not assigned to the inside by the most useful interior heuristics. As shown in Table 18, positions displaying this pattern of variation generally lie inside the folded structure.

Table 15

Splits as interior predictors as a function of PAM width of the subfamilies and different amino acid types

	AAT		ADH		LDH		MYO		PLA		PLC		SOD	
	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
A. *Non-hydrophilic splits*														
							FAMILYVW and CHQST							
000+	89·2	13·6	80·9	16·6	75·8	14·0	87·5	12·0	80·0	7·8	50·0	5·7	70·0	10·2
010+	100·0	11·4	82·6	9·3	61·5	5·0	80·0	6·8	75·0	5·8	100·0	5·7	75·0	8·8
020+	100·0	9·2	83·3	7·3	66·6	5·0	75·0	5·1	100·0	5·8			100·0	7·3
040+	100·0	4·9	90·0	4·4	80·0	5·0	100·0	1·7	100·0	5·8			100·0	2·9
060+	100·0	4·9	100·0	1·9	100·0	3·1			100·0	3·9			100·0	1·4
080+	100·0	2·1	100·0	0·4	100·0	0·6			100·0	1·9			100·0	1·4
100+					100·0	0·6			100·0	1·9				
							FAMILYVW, CHQST, and PG							
000+	54·5	3·2	65·0	6·3	75·0	3·8	40·0	6·8	0·0	0·0	0·0	0·0	50·0	4·4
010+	71·4	2·7	83·3	4·9	100·0	0·6	100·0	5·1	0·0	0·0	0·0	0·0	60·0	4·4
020+	71·4	2·7	71·4	2·4	100·0	0·6	100·0	5·1					60·0	4·4
040+	100·0	1·6	66·6	1·9	100·0	0·6	100·0	1·7					60·0	4·4
060+	100·0	1·6	100·0	0·9	100·0	0·6							60·0	4·4
080+	100·0	1·0											100·0	1·4
100+														
B. *Neutral splits*														
							CHQST							
000+	55·5	5·4	80·0	5·8	50·0	3·1	50·0	1·7	80·0	15·6	100·0	11·4	100·0	8·8
010+	52·9	4·9	85·7	5·8	40·0	1·2	50·0	1·7	80·0	15·6	100·0	11·4	100·0	8·8
020+	52·9	4·9	84·6	5·3	40·0	1·2	50·0	1·7	80·0	15·6	100·0	11·4	100·0	8·8
040+	50·0	4·3	91·6	5·3	40·0	1·2	50·0	1·7	77·7	13·7	100·0	11·4	100·0	8·8
060+	50·0	4·3	91·6	5·3	40·0	1·2	50·0	1·7	77·7	13·7	100·0	11·4	100·0	8·8
080+	53·3	4·3	90·9	4·9	40·0	1·2	50·0	1·7	77·7	13·7	100·0	11·4	100·0	8·8
100+	61·5	4·3	90·9	4·9	40·0	1·2	50·0	1·7	77·7	13·7	100·0	11·4	100·0	8·8
120+			90·9	4·9	40·0	1·2	50·0	1·7	77·7	13·7			100·0	8·8
140+			90·9	4·9			50·0	1·7	77·7	13·7			100·0	8·8
160+			90·9	4·9			50·0	1·7					100·0	8·8
180+			90·9	4·9			50·0	1·7						
							CHQST and PG							
000+	50·0	2·1	75·0	1·4	100·0	1·2	0·0	0·0	60·0	4·4				
010+	60·0	1·6	75·0	1·4	100·0	0·6	0·0	0·0	60·0	4·4				
020+	66·6	1·0	75·0	1·4	100·0	0·6	0·0	0·0	50·0	2·9				
040+	50·0	0·5	100·0	1·4	100·0	0·6	0·0	0·0	66·6	2·9				
060+	50·0	0·5	100·0	0·9	100·0	0·6	0·0	0·0	66·6	2·9				
080+	0·0	0·0	0·0	0·0					100·0	2·9				

C. *Amphiphilic splits*														
FAMILYVW and KREND														
000+	35·7	2·7	66·6	6·8	33·3	1·9	33·3	5·17	50·0	5·8	30·0	8·5	60·0	4·4
010+	30·0	1·6	69·2	4·4	50·0	0·6	0·0	0·0	50·0	5·8	12·5	2·8	60·0	4·4
020+	30·0	1·6	66·6	2·9	50·0	0·6	0·0	0·0	66·6	3·9	20·0	2·8	66·6	2·9
040+	33·3	0·5	75·0	1·4	50·0	0·6	0·0	0·0	100·0	1·9	20·0	2·8	66·6	2·9
060+	33·3	0·5	100·0	0·9	50·0	0·6	0·0	0·0			0·0	0·0	100·0	2·9
080+			100·0	0·9	50·0	0·6	0·0	0·0					0·0	0·0
100+			100·0	0·4	50·0	0·6	0·0	0·0					0·0	0·0
120+			100·0	0·4			0·0	0·0						
140+							0·0	0·0						
160+							0·0	0·0						
FAMILYVW, KREND and PG														
000+	0·0	0·0	33·3	1·9	0·0	0·00	14·2	1·7	100·0	1·9	0·0	0·0	25·0	2·9
010+	0·0	0·0	22·2	0·9			25·0	1·7					16·6	1·4
020+	0·0	0·0	16·6	0·4			0·0	0·0					33·3	1·4
040+	0·0	0·0	33·3	0·4									50·0	1·4
060+	0·0	0·0	33·3	0·4									0·0	0·0
080+	0·0	0·0	100·0	0·4										
FAMILYVW, KREND and CHQST														
000+	21·0	6·5	35·5	10·2	30·7	5·0	9·5	3·4	24·0	11·7	6·2	2·8	15·7	4·4
010+	15·6	2·7	52·1	5·8	100·0	1·2	20·0	1·7	40·0	7·8	100·0	2·8	20·0	2·9
020+	20·0	2·7	50·0	1·9	100·0	1·2	50·0	1·7	33·3	1·9	100·0	2·8	50·0	2·9
040+	22·2	1·0	50·0	0·4	100·0	0·6	100·0	1·7	0·0	0·0	100·0	2·8	66·6	2·9
060+	22·2	1·0	100·0	0·4	100·0	0·6	100·0	1·7					66·6	2·9
080+	66·6	1·0											100·0	1·4
FAMILYVW, KREND, CHQST and PG														
000+	8·3	1·0	19·0	3·9	14·2	0·6	13·6	5·1	21·6	15·6	0·0	0·0	8·0	2·9
010+	11·1	0·5	40·0	1·9	0·0	0·0	20·0	1·7	33·3	9·8	0·0	0·0	12·5	2·9
020+	20·0	0·5	50·0	0·9	0·0	0·0	0·0	0·0	50·0	1·9			25·0	1·4
040+	0·0	0·0	50·0	0·4					100·0	1·9			100·0	1·4
060+	0·0	0·0							100·0	1·9				
D. *Non-hydrophobic splits*														
KREND and CHQST														
000+	16·6	3·2	12·5	1·4	23·8	3·1	21·4	5·1	60·0	5·8	0·0	0·0	37·5	4·4
010+	20·8	2·7	9·0	0·4	0·0	0·0	33·3	5·1	75·0	5·8	0·0	0·0	37·5	4·4
020+	21·0	2·1	14·2	0·4	0·0	0·0	66·6	3·4	100·0	5·8	0·0	0·0	75·0	4·4
040+	16·6	1·0	0·0	0·0	0·0	0·0	100·0	1·7	100·0	3·9			00·0	4·4
060+	16·6	1·0	0·0	0·0					100·0	1·9			00·0	4·4
080+	25·0	0·5							0·0	0·0			00·0	2·9
100+									0·0	0·0				
KREND, CHQST, and PG														
000+	11·1	0·5	37·5	2·9	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	20·0	1·4
010+	20·0	0·5	40·0	1·9			0·0	0·0	0·0	0·0	0·0	0·0	25·0	1·4
020+			37·5	1·4			0·0	0·0	0·0	0·0			50·0	1·4
040+			60·0	1·4									50·0	1·4
060+			66·6	0·9									50·0	1·4
080+			100·0	0·4									0·0	0·0

Table 15 *continued*

	AAT		ADH		LDH		MYO		PLA		PLC		SOD	
	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
E. *Hydrophilic Splits*														
							KREND							
000+	41·3	6·5	46·1	2·9	28·5	3·8	0·0	0·0	50·0	3·9	50·0	8·5	100·0	5·8
010+	46·1	6·5	75·0	2·9	33·3	3·8	0·0	0·0	66·6	3·9	60·0	8·5	100·0	5·8
020+	50·0	6·5	85·7	2·9	33·3	3·1	0·0	0·0	100·0	3·9	60·0	8·5	100·0	5·8
040+	47·8	6·0	85·7	2·9	33·3	3·1	0·0	0·0	100·0	3·9	60·0	8·5	100·0	5·8
060+	47·8	6·0	85·7	2·9	33·3	3·1	0·0	0·0	100·0	3·9	60·0	8·5	100·0	5·8
080+	50·0	5·4	83·3	2·4	38·4	3·1	0·0	0·0	100·0	3·9	60·0	8·5	100·0	5·8
100+	50·0	4·9	83·3	2·4	38·4	3·1	0·0	0·0	100·0	3·9	60·0	8·5	100·0	4·4
120+			83·3	2·4	38·4	3·1	0·0	0·0	100·0	3·9			100·0	4·4
140+			100·0	2·4			0·0	0·0	100·0	3·9			100·0	4·4
160+			100·0	2·4			0·0	0·0	100·0	4·41				
180+			100·0	2·4			0·0	0·0						
							KREND and PG							
000+	11·1	0·54	25·0	0·98	33·3	0·63	0·0	0·0	0·0	0·0	0·0	0·0		
010+	0·0	0·0	33·3	0·98	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0		
020+	0·0	0·0	50·0	0·98	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0		
040+	0·0	0·0	66·6	0·98	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0		
060+	0·0	0·0	66·6	0·98	33·3	3·44	0·0	0·0	0·0	0·0	0·0	0·0		
080+	0·0	0·0	100·0	0·49	40·0	3·44					0·0	0·0		
100+			100·0	0·49	25·0	1·72					0·0	0·0		
120+			100·0	0·49	25·0	1·72								
140+					33·3	1·72								
160+					50·0	1·72								
180+					100·0	1·72								
F. *Parsing splits*														
							PG							
000+	75·0	8·19	68·7	5·39	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	63·6	10·29
010+	75·0	8·19	68·7	5·39	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	63·6	10·29
020+	75·0	8·19	68·7	5·39	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	63·6	10·29
040+	75·0	8·19	68·7	5·39	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	63·6	10·29
060+	75·0	8·19	68·7	5·39	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	63·6	10·29
080+	75·0	8·19	68·7	5·39	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	70·0	10·29
100+	75·0	8·19	64·2	4·41	56·2	5·73	100·0	1·72	100·0	3·92	27·2	8·57	70·0	10·29
120+			64·2	4·41	56·2	5·73	100·0	1·72	100·0	3·92			66·6	8·82
140+			64·2	4·41			100·0	1·72	100·0	3·92			66·6	8·82
160+			64·2	4·41			100·0	1·72					66·6	8·82
180+			64·2	4·41			100·0	1·72						

To be recorded *at a particular row, the split* must be observed in subfamilies having the PAM width indicated in the left column or higher. Different types of split are defined in Table 15. Less than 50% accessibility to a probe of 1·4 Å defines an interior residue. When no position in the alignment fits the criterion, no entry is made. Zero values therefore indicate cases where all of the positions identified by the indicated heuristic are surface positions.

Table 16

Hydrophobic variable positions as interior indicators (interior indicating amino acids = FAMILYVW)

PAM	AAT		ADH		LDH		MYO		PLA		PLC		SOD	
Window	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
A. *One variable subfamily*														
0–010	66·6	7·6	80·9	8·3	76·9	25·4	100·0	12·0	100·0	8·5	100·0	1·4		
0–020	72·0	9·8	85·1	11·2	76·9	25·4	90·9	17·2	100·0	1·9	83·3	14·2	100·0	5·8
0–040	74·1	12·5	84·2	15·6	76·7	27·3	94·1	27·5	75·0	5·8	83·3	14·2	100·0	17·6
0–060	74·1	12·5	86·6	19·1	75·8	28·0	94·4	29·3	80·0	7·8	75·0	25·7	100·0	20·5
0–080	76·7	23·4	88·0	21·5	75·0	28·6	90·9	34·4	80·0	15·6	75·0	25·7	100·0	20·5
0–100	76·4	28·4	88·0	21·5	75·0	28·6	92·5	43·1	72·7	15·6	75·0	25·7	100·0	20·5
0–120	76·4	28·4	88·0	21·5	76·1	30·5	92·5	43·1	72·7	15·6	75·0	25·7	100·0	22·0
0–140	76·4	28·4	88·8	23·5	76·1	30·5	92·5	43·1	72·7	15·6	75·0	25·7	100·0	22·0
0–160	76·4	28·4	88·8	23·5	76·1	30·5	92·5	43·1	72·7	15·6	75·0	25·7	100·0	22·0
0–180	76·4	28·4	88·8	23·5	76·1	30·5	86·6	44·8	72·7	15·6	75·0	25·7	100·0	22·0
0–200	76·4	28·4	88·8	23·5	76·1	30·5	86·6	44·8	72·7	15·6	75·0	25·7	100·0	22·0
B. *Two variable subfamilies*														
0–010	85·7	3·2	100·0	2·9	76·1	10·1	100·0	3·4						
0–020	75·0	3·2	100·0	4·4	76·1	10·1	100·0	5·1	100·0	8·5	100·0	2·9		
0–040	75·0	3·2	86·6	6·3	82·7	15·2	90·0	15·5	100·0	3·9	100·0	8·5	100·0	8·8
0–060	75·0	3·2	88·2	7·3	84·3	17·1	90·0	15·5	100·0	5·8	80·0	11·4	100·0	16·1
0–080	77·2	9·2	90·9	9·8	84·8	17·8	91·6	18·9	100·0	7·8	80·0	11·4	100·0	16·1
0–100	77·2	9·2	91·6	10·7	84·8	17·8	92·3	20·6	100·0	7·8	80·0	11·4	100·0	16·1
0–120	77·2	9·2	91·6	10·7	81·0	19·1	92·3	20·6	100·0	7·8	80·0	11·4	100·0	16·1
0–140	77·2	9·2	88·4	11·2	81·0	19·1	92·3	20·6	100·0	7·8	80·0	11·4	100·0	16·1
0–160	77·2	9·2	88·4	11·2	81·0	19·1	92·3	20·6	100·0	7·8	80·0	11·4	100·0	16·1
0–180	77·2	9·2	88·4	11·2	81·0	19·1	92·3	20·6	100·0	7·8	80·0	11·4	100·0	16·1
0–200	77·2	9·2	88·4	11·2	81·0	19·1	92·3	20·6	100·0	7·8	80·0	11·4	100·0	16·1
C. *Three variable subfamilies*														
0–010	100·0	0·5	91·6	7·0										
0–020	100·0	0·5	100·0	1·4	91·6	7·0	100·0	2·8	100·0	1·4				
0–040	100·0	0·5	100·0	2·9	81·2	8·2	50·0	1·7	100·0	2·8	100·0	1·4		
0–060	100·0	0·5	100·0	2·9	81·2	8·2	50·0	1·7	100·0	3·9	100·0	2·8	100·0	1·4
0–080	100·0	0·5	100·0	3·9	81·2	8·2	66·6	3·4	100·0	3·9	100·0	2·8	100·0	1·4
0–100	100·0	0·5	100·0	3·9	81·2	8·2	66·6	3·4	100·0	3·9	100·0	2·8	100·0	1·4
0–120	100·0	0·5	100·0	3·9	81·2	8·2	75·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4
0–140	100·0	0·5	100·0	3·9	81·2	8·2	75·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4
0–160	100·0	0·5	100·0	3·9	81·2	8·2	75·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4
0–180	100·0	0·5	100·0	3·9	81·2	8·2	75·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4
0–200	100·0	0·5	100·0	3·9	81·2	8·2	75·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4

The maximum PAM width for the cluster of subfamilies where designated number of variable subfamilies is seen is given in the left column. Less than 50% accessibility to a probe of 1·4 Å defines an interior residue. When no position in the alignment fits the criterion, no entry is made. Zero values therefore indicate cases where all of the positions identified by the indicated heuristic are surface positions.

Table 17

Hydrophobic variable positions as interior indicators (interior indicating amino acids = FAMILYVWPG)

PAM	AAT		ADH		LDH		MYO		PLA		PLC		SOD	
Window	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
A. *One variable subfamily*														
0–010	68·1	8·1	85·7	11·7	76·7	27·3	87·5	12·0	100·0	11·4	100·0	2·9		
0–020	70·3	10·3	86·8	16·1	76·7	27·3	84·6	18·9	100·0	3·9	85·7	17·1	100·0	8·8
0–040	68·5	13·1	85·1	22·5	76·6	29·2	89·4	29·3	83·3	9·8	85·7	17·1	100·0	22·0
0–060	68·5	13·1	85·4	25·9	74·6	29·9	90·0	31·0	85·7	11·7	78·5	31·4	100·0	27·9
0–080	71·2	25·6	85·2	28·4	73·8	30·5	87·5	36·2	83·3	19·6	78·5	31·4	95·0	27·9
0–100	71·7	30·6	84·0	28·4	73·8	30·5	89·6	44·8	76·9	19·6	78·5	31·4	95·0	27·9
0–120	71·7	30·6	84·0	28·4	72·2	33·1	89·6	44·8	76·9	19·6	78·5	31·4	95·4	30·8
0–140	71·7	30·6	84·4	31·8	72·2	33·1	89·6	44·8	76·9	19·6	78·5	31·4	95·4	30·8
0–160	71·7	30·6	84·4	31·8	72·2	33·1	89·6	44·8	76·9	19·6	78·5	31·4	95·4	30·8
0–180	71·7	30·6	84·4	31·8	72·2	33·1	84·3	46·5	76·9	19·6	78·5	31·4	95·4	30·8
0–200	71·7	30·6	84·4	31·8	72·2	33·1	84·3	46·5	76·9	19·6	78·5	31·4	95·4	30·8
B. *Two variable subfamilies*														
0–010	85·7	3·2	100·0	3·4	77·2	10·8	66·6	3·4						
0–020	75·0	3·2	100·0	5·8	77·2	10·8	75·0	5·1	100·0	8·5	100·0	4·4		
0–040	75·0	3·2	87·5	10·2	80·6	15·9	83·3	17·2	100·0	3·9	100·0	8·5	100·0	10·2
0–060	75·0	3·2	88·4	11·2	82·3	17·8	83·3	17·2	100·0	5·8	80·0	11·4	100·0	19·1
0–080	77·2	9·2	90·6	14·2	82·8	18·4	85·7	20·6	100·0	11·7	80·0	11·4	100·0	20·5
0–100	77·2	9·2	91·1	15·1	82·8	18·4	86·6	22·4	100·0	11·7	80·0	11·4	100·0	20·5
0–120	77·2	9·2	91·1	15·1	79·4	19·7	86·6	22·4	100·0	11·7	80·0	11·4	100·0	20·5
0–140	77·2	9·2	89·4	16·6	79·4	19·7	86·6	22·4	100·0	11·7	80·0	11·4	100·0	20·5
0–160	77·2	9·2	89·4	16·6	79·4	19·7	86·6	22·4	100·0	11·7	80·0	11·4	100·0	20·5
0–180	77·2	9·2	89·4	16·6	79·4	19·7	86·6	22·4	100·0	11·7	80·0	11·4	100·0	20·5
0–200	77·2	9·2	89·4	16·6	79·4	19·7	86·6	22·4	100·0	11·7	80·0	11·4	100·0	20·5
C. *Three variable subfamilies*														
0–010	100·0	0·5	91·6	7·0	0·0	0·0								
0–020	100·0	0·5	100·0	1·9	91·6	7·0	0·0	0·0	100·0	2·8	100·0	1·4		
0–040	100·0	0·5	100·0	3·4	81·2	8·2	50·0	3·4	100·0	2·8	100·0	1·4		
0–060	100·0	0·5	87·5	3·4	81·2	8·2	50·0	3·4	100·0	3·9	100·0	2·8	100·0	1·4
0–080	100·0	0·5	90·0	4·4	81·2	8·2	60·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4
0–100	100·0	0·5	90·0	4·4	81·2	8·2	60·0	5·1	100·0	3·9	100·0	2·8	100·0	1·4
0–120	100·0	0·5	90·0	4·4	81·2	8·2	66·6	6·8	100·0	3·9	100·0	2·8	100·0	1·4
0–140	100·0	0·5	90·0	4·4	81·2	8·2	66·6	6·8	100·0	3·9	100·0	2·8	100·0	1·4
0–160	100·0	0·5	90·0	4·4	81·2	8·2	66·6	6·8	100·0	3·9	100·0	2·8	100·0	1·4
0–180	100·0	0·5	90·0	4·4	81·2	8·2	66·6	6·8	100·0	3·9	100·0	2·8	100·0	1·4
0–200	100·0	0·5	90·0	4·4	81·2	8·2	66·6	6·8	100·0	3·9	100·0	2·8	100·0	1·4

The maximum PAM width for the cluster of subfamilies where designated number of variable subfamilies is seen is given in the left column. Less than 50% accessibility to a probe of 1·4 Å defines an interior residue. When no position in the alignment fits the criterion, no entry is made. Zero values therefore indicate cases where all of the positions identified by the indicated heuristic are surface positions.

Table 18

Variable positions containing CHQST but no KREND as indicators of an interior position

PAM	AAT		ADH		LDH		MYO		PLA		PLC		SOD	
Window	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
A. *One variable subfamily*														
0–010	35·0	3·8	46·8	7·3	49·1	17·8	11·7	3·4	25·0	3·9	0·0	0·0	0·0	0·0
0–020	40·7	6·0	44·2	13·2	47·4	17·8	20·8	8·6	31·5	11·7	0·0	4·4	0·0	10·7
0–040	30·7	6·5	40·2	17·1	47·7	20·3	25·0	13·7	28·1	17·6	0·0	4·4	0·0	8·8
0–060	30·7	6·5	43·1	21·5	46·3	20·3	25·7	15·5	24·3	19·6	0·0	8·8	0·0	15·7
0–080	38·3	12·5	48·4	29·9	47·2	22·2	27·6	22·4	25·5	23·5	0·0	11·7	0·0	18·6
0–100	46·9	20·7	49·6	31·8	47·2	22·2	27·6	22·4	25·0	23·5	0·0	13·2	0·0	20·4
0–120	46·9	20·7	49·6	31·8	50·6	26·7	27·6	22·4	26·5	25·4	0·0	22·0	0·0	28·3
0–140	46·9	20·7	52·1	35·2	50·6	26·7	27·6	22·4	26·5	25·4	0·0	22·0	0·0	28·3
0–160	46·9	20·7	52·1	35·2	50·6	26·7	27·6	22·4	26·5	25·4	0·0	22·0	0·0	28·3
0–180	46·9	20·7	52·1	35·2	50·6	26·7	30·6	25·8	26·5	25·4	0·0	22·0	0·0	28·3
0–200	46·9	20·7	52·1	35·2	50·6	26·7	30·6	25·8	26·5	25·4	0·0	22·0	0·0	28·3
B. *Two variable subfamilies*														
0–010	25·0	0·5	75·0	1·4	30·0	1·9	12·5	1·7	0·0	0·0				
0–020	20·0	0·5	50·0	3·9	36·3	2·5	18·1	3·4	0·0	0·0	0·0	0·0	0·0	0·0
0–040	28·5	1·0	47·6	4·9	50·0	5·0	31·5	10·3	0·0	0·0	0·0	0·0	0·0	0·0
0–060	28·5	1·0	47·8	5·3	50·0	5·0	31·5	10·3	0·0	0·0	0·0	0·0	0·0	0·0
0–080	30·0	1·6	48·0	5·8	47·0	5·0	30·4	12·0	10·0	3·9	0·0	1·4	0·0	11·1
0–100	30·0	1·6	51·7	7·3	47·0	5·0	33·3	13·7	13·6	5·8	0·0	1·4	0·0	11·1
0–120	30·0	1·6	51·7	7·3	52·6	6·3	33·3	13·7	13·6	5·8	0·0	1·4	0·0	11·1
0–140	30·0	1·6	54·8	8·3	52·6	6·3	33·3	13·7	13·6	5·8	0·0	1·4	0·0	11·1
0–160	30·0	1·6	54·8	8·3	52·6	6·3	33·3	13·7	13·6	5·8	0·0	1·4	0·0	11·1
0–180	30·0	1·6	54·8	8·3	52·6	6·3	36·0	15·5	13·6	5·8	0·0	1·4	0·0	11·1
0–200	30·0	1·6	54·8	8·3	52·6	6·3	36·0	15·5	13·6	5·8	0·0	1·4	0·0	11·1
C. *Three variable subfamilies*														
0–010	0·0	0·0	0·0	0·0										
0–020	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0				
0–040	0·0	0·0	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–060	0·0	0·0	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–080	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–100	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–120	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–140	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–160	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–180	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		
0–200	60·0	1·4	25·0	0·6	25·0	3·4	0·0	0·0	0·0	0·0	0·0	0·0		

The maximum PAM width for the cluster of subfamilies where designated number of variable subfamilies is seen is given in the left column. Less than 50% accessibility to a probe of 1·4 Å defines an interior residue. When no position in the alignment fits the criterion, no entry is made. Zero values therefore indicate cases where all of the positions identified by the indicated heuristic are surface positions.

Because the variation involves an internally placed functional group (e.g. -OH groups) that often interacts with other hydrogen-bonding side-chains from amino acids elsewhere in the polypeptide chain, these positions often undergo mutation that is compensated by mutations elsewhere in the protein chain. Thus, they can be useful in covariation analysis schemes that detect contacts between distant parts of the polypeptide chain (Benner & Gerloff, 1991).

4. Discussion

Surface and interior heuristics based on patterns of conservation and variation within a set of aligned homologous protein sequences offer predictions that are typically between 80 and 100% accurate; the best heuristics applicable to any individual protein family generally make predictions with >90% accuracy. Although such high levels of accuracy make them useful, the heuristics make errors, and it is instructive to ask why.

One source of error undoubtedly arises from the breakdown of the assumption central to any method that builds a conformational model from a set of aligned homologous sequences: that homologous proteins have generally similar folded structures (Chothia & Lesk, 1986; Summers *et al.*, 1987). Conformation does in fact diverge along with divergence in sequence. Therefore, the exposure of the side-chains of homologous amino acids in two homologous proteins need not be the same. Indeed, at greater evolutionary divergence, entire secondary structural units can be gained or lost. Clearly, in cases where amino acids matched in an alignment have different surface accessibilities in different proteins in the same family, any binary assignment (e.g. surface or interior) must be correct for some and incorrect for others. Whether an assignment for these positions is counted as an error or not depends fortuitously on which branch of the family is represented in the crystal structure used to score the assignments.

Next, misalignments are a potent source of errors. Misalignments are a failure to match homologous amino acids (amino acids encoded by codons that are descendants of a common ancestral codon). These generally cause surface heuristics to make overpredictions and interior heuristics to have lower coverages. They are generally avoided by including in a multiple alignment only proteins that have diverged by less than 200 PAM units. However, above 150 PAM, multiple alignments invariably contain some poor segments. In practice, marginal regions of an alignment are normally recognizable; the confidence attributed to surface predictions should be reduced accordingly by the biochemist evaluating the prediction. Alternatively, surface heuristics may be applied to subfamilies of the alignment separately, where sequences in each subfamily are reliably aligned.

Third, interior heuristics often make errors when applied to families of proteins that physiologically aggregate as multimers (dimers and tetramers). Residues on the surface of the subunit involved in the formation of quaternary contacts often behave as interior residues during divergent evolution, and are often assigned to the inside by interior heuristics. To guide site-directed mutagenesis or to search for antigenic sites (Hopp & Woods, 1981; Holbrook *et al.*, 1990), such residues should perhaps not be considered as errors, as the positions are indeed not exposed to solvent in the native quaternary structure. However, as surface exposures are calculated for individual subunits alone, and as tertiary structure predictions build models for individual subunits, these assignments are counted as errors in the results reported here.

Accordingly, interior heuristics are most inaccurate with proteins (lactate dehydrogenase, superoxide dismutase) that have retained a dimeric structure throughout divergent evolution. Accuracies of interior heuristics are higher with monomeric proteins (myoglobin, phospholipase and plastocyanin), or proteins whose quaternary structure has diverged during divergent evolution (alcohol dehydrogenase). The lesser accuracy of the interior assignments for aspartate aminotransferase can be explained by the dimeric quaternary structure of the protein, the small number of sequences in the alignment, and the relatively small evolutionary distance separating the sequences in the alignment. The first fact implies that some residues on the surface of the subunit are buried in the native protein's quaternary structure. The modest overall sequence divergence within the family implies that the conformational significance of conservation is diminished. The lower levels of accuracy observed with PLA may be ascribed to its contact with membranes; membrane-bound and transmembrane proteins appear to display hydrophobic variability on the faces in contact with the lipid (Rees *et al.*, 1989).

The alcohol dehydrogenase (ADH) family is especially interesting. Coverage by the first heuristic (Table 8) is remarkably low and the level of accuracy remarkably high, despite the fact that the protein exists physiologically either as a dimer (e.g. in horse liver) or as a tetramer (in yeast). The patterns of variation were used to predict the quaternary structure of yeast alcohol dehydrogenase, for which no crystal structure is available (Benner, 1989). The model proposes contacts between the subunits in the yeast tetramer that are different from those in the dimeric protein from horse liver, where a crystal structure has been solved (Eklund *et al.*, 1976). Although the quaternary structural model predicted for yeast alcohol dehydrogenase has not been confirmed by crystallographic work, the model underlies a program of site-directed mutagenesis in these laboratories (Weinhold *et al.*, 1991).

In practice, the quaternary structure of a protein family is generally known from biochemical studies. If the protein forms a multimer, remedial steps must be taken in using the interior assignments to assemble secondary and tertiary structure (Benner, 1989). This underscores a more general point. As with conformational analysis in other branches of chemistry, structure prediction in protein chemistry is best done by those who understand the reactivity, biochemistry and biology of the protein family, and use what they know in making predictions (Benner, 1989; Benner & Ellington, 1990; Benner & Gerloff, 1991; Benner 1992*a*).

The accuracy and coverage obtained with a heuristic depend, often strongly, on the number of proteins in a multiple alignment and their distribution on an evolutionary tree. This point has important implications with respect to the strategy used to develop methods for structure prediction. It is common within the field to average prediction heuristics over large numbers of protein families, and to report only averaged scores. Indeed, some have come to insist that this is the only proper way to evaluate prediction heuristics, assuming that prediction is fundamentally statistical in nature (Robson & Garnier, 1993). Table 5 shows that this approach loses most of the interesting information. As in chemistry generally, conformational analysis is not fundamentally statistical; individual molecules have individual properties, these properties are ultimately what interest the biochemist, and uninhibited aggregation of these properties undermines efforts to develop understanding. While the problem of obtaining statistically representative samples remains, other ways to solve it must be sought.

Further, when collecting sequence data for making a prediction of structure, sequences should be obtained experimentally to balance and fill out the evolutionary tree. An estimate of the rate of divergence of a protein sequence and an understanding of the organismal evolutionary tree (based, for example, on ribosomal RNA sequences, Pace *et al.*, 1986) are together generally adequate to allow the selection of organisms to provide sequences that do this.

The dependence of accuracy and coverage on the number and evolutionary diversity of a protein family creates practical problems, however, especially if experimental work is not possible to collect sequences additional to those found in the database. For both surface and interior predictions, a large number of heuristics exist each having particular prescriptions for the number of variable or conserved subfamilies, the PAM width used to construct these subfamilies, the sets of amino acids that are considered surface-indicating and interior-indicating, the distribution of these amino acids within the subfamilies, and the pattern of variation and conservation within the subfamilies. It has been relatively easy to write a computer program that applies all of these heuristics to a specific alignment. It is less easy to present the resulting output (surface, interior, parsing and active site assignments) in a form that helps the biochemist build a structural model of the protein or guide experimental work.

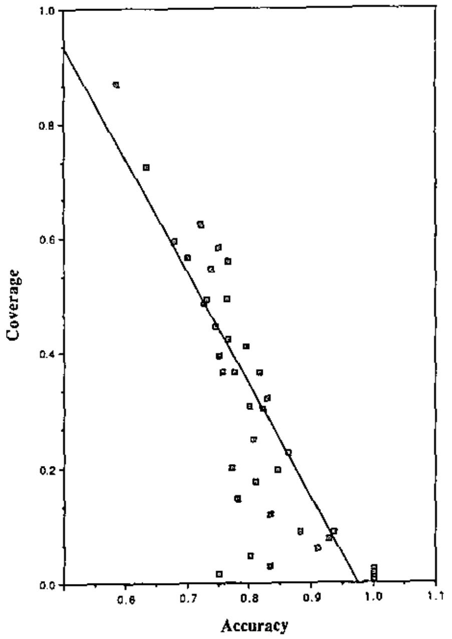


Figure 6. Inverse relation between coverage and accuracy for the surface predictions made by a set of surface heuristics applied to the alcohol dehydrogenase family. Surface-indicating amino acids are KRENDQST.

One solution to this problem is based on the generalization that the number of assignments made by a particular heuristic within a class of analogous heuristics is inversely related to the accuracy of the heuristic (Fig. 6). To the extent that this relationship holds, the strongest assignment at any particular position is made by the heuristic that makes the lowest number of assignments in the alignment as a whole. Thus, as presently implemented, the biochemist is presented with at most a single surface assignment and a single interior assignment for each position in the multiple alignment. The surface assignment reported is that made by the heuristic that (1) assigns the designated position to the surface and (2) makes the smallest number of surface assignments in the alignment overall. The interior assignment reported is the one made by the interior heuristic that assigns (1) the designated position to the interior and (2) makes the smallest number of interior assignments in the alignment overall. The remaining assignments, presumably weaker because they are made by heuristics that assign more positions in the alignment, are not reported. The output for each position is a description of the heuristic that makes the assignments together with the fraction of the alignment overall that is assigned by this heuristic.

The surface and interior assignments made by the heuristics evaluated here have considerable value, both in their own right (Holbrook *et al.*, 1990) and as starting points for the prediction of secondary structure. In their simplest application, 3·6 residue periodicity in surface and interior assignments have been used to indicate a surface helix in proteins with unknown structures, while segments of consecutive interior residues (up to 8 amino acid residues in length) have been used to predict interior beta strands.

A systematic evaluation of heuristics for predicting secondary structure based on surface and interior assignments will be the topic of the next paper in this series. Nevertheless, it is worth noting here that nearly a dozen *bona fide* predictions have now been made using a formalism based on the surface and interior heuristics evaluated in this paper. Of these, two first stage unrefined predictions (for the SH3 domain and the MoFe nitrogenase protein; Benner *et al.*, 1993*a*; Gerloff *et al.*, 1993*a*) and two refined predictions (for protein kinase (Benner & Gerloff, 1991) and for the hemorrhagic metalloproteinases (Gerloff *et al.*, 1993*b*) can now be examined in light of subsequently determined crystal and NMR structures (Knighton *et al.*, 1991; Musacchio *et al.*, 1992*b*; Yu *et al.*, 1992; Kim & Rees, 1992; Koyama *et al.*, 1993).

Independent evaluation of the unrefined predictions suggests that they produce per-residue three state scores similar to those obtained using classical methods averaged over a set of aligned homologous sequences (Rost & Sander, 1992), but perform better in assigning core segments (Thornton *et al.*, 1991). In part, this is because the method identifies regions in the protein sequence that do not form part of the core fold, and disregards these. Errors are concentrated in regions where the multiple alignment is poor, where conformation has diverged in the protein family, and near the active site, where functional constraints important for catalysis obscure those that indicate secondary structure. For example, in the MoFe nitrogenase prediction (Gerloff *et al.*, 1993*a*), ten helices were predicted; all were later shown to correspond to a helix in the experimental structure. Near the active site and in regions where the alignment was poor, however, the prediction was less satisfactory.

Nevertheless, the approach remains controversial. For example, the refined prediction for protein kinase has been viewed as "remarkably accurate" and a "spectacular achievement" (Knighton *et al.*, 1991; Lesk & Boswell, 1992). Others, however, have regarded these comments as "exaggerated" (Rost *et al.*, 1993). Regardless of which view is correct, the quality of the secondary structure prediction for protein kinase in the core of the first domain (together with covariation analysis and active site assignments) was adequate to allow Benner & Gerloff (1991) to guess correctly that the core was built from antiparallel beta strands. This fold was unusual among kinases (Knighton *et al.*, 1991), and contrasted sharply with structure predictions made in other laboratories (Bramson *et al.*, 1984; Shoji *et al.*, 1983; Sternberg & Taylor, 1984; Taylor *et al.*, 1988; Fry *et al.*, 1986) using Chou-Fasman analyses (Chou & Fasman, 1978), GOR analyses (Garnier *et al.*, 1978), or analyses based on sequence motifs and consensus sequence elements (Dever *et al.*, 1987; Bairoch, 1991). In protein kinase, all predictions based on classical methods proved to be less satisfactory (Bork, 1992). A detailed review of the prediction for the protein kinase family in light of the crystal structure is provided by Benner (1992*a*).

5. Glossary

Accuracy of a set of assignments: The number of correct assignments divided by the total number of assignments made, expressed in percentage.

Alignment anchor: A position in an alignment that is sufficiently conserved across the entire alignment, or undergoes only conservative substitution, such that homologous amino acids in different proteins are reliably matched in the alignment.

Amphiphilic split in a cluster of subfamilies with MPW of X: Designates a position in an alignment where none of the subfamilies in the cluster of subfamilies at MPW = X is variable, where at least one subfamily contains a hydrophobic residue, and at least one subfamily contains a hydrophilic residue.

Amino acid: Used to designate the amino acids as abstractions (compare "residue").

Amino acid type: The 20 proteinogenic amino acids may be divided into types based on a specified property (e.g. hydrophobic, structure-disrupting, surface-indicating).

APC (all proteins conserved): Designates a position in an alignment where all proteins being considered have the same amino acid at that position.

Cluster of subfamilies with a maximum MPW = X: The proteins in the alignment are divided into subfamilies with different overall levels of PAM width.

CMX Y (Count minus X): Designates a position in the alignment where all but X proteins have amino acid Y.

Coverage of a prediction: The number of positions in an alignment correctly assigned a particular structural attribute (surface positions, interior positions, etc.) divided by the number of positions in the study object that display this attribute.

Distributed parse: A parse built from parsing elements that appear in different subfamilies of the alignment at neighboring position numbers.

Functional subfamilies: Members of each functional subfamily (or subfamily) share among themselves, and are distinct from members of other functional subfamilies, a particular biological function or catalytic behavior.

Helix plot: A projection of a helix down its long axis, showing the directions that side-chains at different positions along the helix protrude.

Hydrophilic: In this article, hydrophilic is used to indicate an experimentally measured property of an amino acid side-chain. The experimental method can, of course, vary. Typical hydrophilic amino acids are KREND.

Hydrophilic split in a cluster of subfamilies with MPW of X: Designates a position in an alignment where each subfamily displays no variation, and all amino acids in each subfamily are surface-indicating.

Hydrophobic: In this article, hydrophobic is used to indicate an experimentally measured property of an amino acid side-chain. The experimental method can, of course, vary. Typical hydrophobic amino acids are FAMILYVW.

Hydrophobic anchor for an external loop: A position with a hydrophobic amino acid in most proteins appearing in a segment that is a parse or otherwise assigned as a surface loop.

Hydrophobic split in a cluster of subfamilies at a specified MaxPW: Designates a position in an alignment where none of the subfamilies in the cluster of subfamilies at a specified MaxPW is variable, and the residue types are interior-indicating.

Hydrophobic variable in a cluster of subfamilies at a specified MaxPW: Designates a position in an alignment where at least one of the subfamilies in the cluster of subfamilies at the specified MaxPW is variable, but where none of the proteins has KREND or CHQST.

Indifferent residue: One of the following amino acids: Cys, His, Gln, Ser, Thr, Gly and Pro.

Inside arc: The side of a helix wheel from which protrude side-chains of residues assigned to the inside of the folded protein structure.

Interior-indicating: Amino acids whose presence at a position in an alignment fulfil one of the criteria for assigning the position to inside of the folded structure of the protein. "Interior-indicating" is a property defined for individual heuristics, in contrast with "hydrophobic", which is defined by an experimental operation.

MaxPW or MPW: Maximum PAM width, the PAM distance of the two most distant proteins within a single subfamily of proteins.

Neutral split in a cluster of subfamilies with MaxPW of X: Designates a position in an alignment where none of the subfamilies in the cluster of subfamilies at MaxPW = X is variable, where every subfamily contains an indifferent residue.

Non-hydrophilic variable in a cluster of subfamilies at a specified MaxPW: Designates a position in an alignment where at least one of the subfamilies in the cluster of subfamilies at the specified MaxPW is variable, but where none of the subfamilies has a KREND, and at least one subfamily has a CHQST.

Non-standard secondary structure: All secondary structures other than an alpha helix or a beta strand (e.g. a 3_{10} helix or a collagen helix).

PAM (accepted point mutations) distance: A measure of the evolutionary distance between two proteins, where the PAM distance is the most probable number of accepted point mutations separating the two sequences per 100 amino acid residues, corresponding to the number of times the first sequence must be transformed by a 1% mutation matrix (a mutation matrix where the sum of all off-diagonal elements is such that a single transformation of a sequence by this matrix yields a protein with 1 mutation per 100 amino acids) to yield the second protein with the highest probability.

PAM width (PW): The PAM width for a set of protein sequences is the PAM distance separating the two most divergent proteins in the subfamily.

Parse: A region of the alignment that divides the alignment into segments whose secondary structure is considered independently.

Parsing string: A sequences of consecutive amino acids in a single protein that indicates that the segment lies between standard secondary structural units (e.g. GG, PP, PG, GP, NN, NS, etc.).

PW = PAM width.

Reflexivity: Designates a position in an alignment where patterns of variation among the proteins being examined suggests a tree-like relationship between these proteins that is different from the evolutionary tree derived from examination of the entire sequences of the proteins. Most commonly, a position displays reflexivity when the pattern of variation involving particular amino acids is the same in two distant subfamilies.

Residue: In this article, a residue is a specific amino acid at a specific position in a polypeptide chain.

Residue type: There are 20 different residue types, corresponding to each of the 20 proteinogenic amino acids.

Split at MaxPW = X: Designates a position in an alignment where none of the subfamilies in the cluster of subfamilies at MaxPW = X is variable.

Standard secondary structural elements: an alpha helix or a beta strand.

String: A set of consecutive positions in the alignment.

Subfamily with a specified MaxPW: A subset of the proteins in an alignment where every sequence in the subfamily is connected to at least one other sequence in the same subfamily by a bridge at or below the specified PAM distance.

Surface arc: The side of a helix wheel from which protrude side-chains of residues assigned to the surface of the protein.

Surface-indicating: Amino acids whose presence at a position in an alignment fulfil one of the criteria for assigning the position to the surface of the folded structure of the protein. "Surface-indicating" is a property defined for individual heuristics, in contrast with "hydrophilic", which is defined by an experimental operation.

Variable subfamily: At a position in an alignment, a subfamily of proteins in the alignment where more than one residue type is present.

We are indebted to Professor Gaston Gonnet and Lukas Knecht for many helpful discussions concerning computational aspects of the work, to Adam Feigin for help setting up the computational facilities, to Wolfgang Weck for providing the PhylTree program, and to Sandoz AG for partial support of this work. M.A.C. was supported by a Wellcome Trust Travelling Fellowship.

References

Allemann, R. K. (1989). Evolutionary guidance as a tool in organic chemistry. Dissertation, no. 8804, Eidgenössiche Technische Hochschule, Zurich.

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **16**, 2241–2245.

Bazan, J. F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci., U.S.A.*, **87**, 6934–6938.

Bazan, J. F. (1992). Unraveling the structure of IL-2. *Science*, **257**, 410–412.

Benner, S. A. (1989). Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Advan. Enzyme Regulat.* **28**, 219–236.

Benner, S. A. (1992*a*). Predicting *de novo* the folded structure of proteins. *Curr. Opin. Struct. Biol.* **2**, 402–412.

Benner, S. A. (1992*b*). Predicting folded structures of proteins. U.S. Patent Application 07/857,224, March 25, 1992.

Benner, S. A. & Ellington, A. D. (1990). Evolution and structural theory: the frontier between chemistry and biochemistry. *Bioorg. Chem. Frontiers*, **1**, 1–70.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: the catalytic domain of protein kinases. *Advan. Enzyme Regulat.* **31,** 121–181.

Benner, S. A., Cohen, M. A., Gonnet, G. H., Berkowitz, D. B. & Johnsson, K. (1992). Reading the palimpsest: contemporary biochemical data and the RNA world. In *The RNA World* (Gesteland, R. & Atkins, J. eds), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, in the press.

Benner, S. A., Cohen, M. A. & Gerloff, D. (1993*a*). A predicted structure for the Src homology 3 domain. *J. Mol. Biol.* **229**, 295–305.

Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1993*b*). Empirical and structural models for insertions and deletions in the evolution of proteins. *J. Mol. Biol.* **229**, 1065–1082.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, U. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature (London)*, **326**, 347–352.

Bork, P. (1992). Mobile modules and motifs. *Curr. Opin. Struct. Biol.* **2**, 413–421.

Bowie, J. U., Luethy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Bramson, H. N., Kaiser, E. T. & Milvan, A. S. (1984). Mechanistic studies of cAMP-dependent protein kinase action. *CRC Crit. Rev. Biochem.* **15**, 93–124.

Brandhuber, B. J., Boone, T., Kenney, W. C. & McKay, D. B. (1987). Three-dimensional structure of interleukin-2. *Science*, **238**, 1707–1709.

Chakravarty, P. K., Mathur, K. B. & Dhar, M. M. (1973). The synthesis of a decapeptide with glycosidase activity. *Experientia*, **29**, 786–788.

Chothia, C. & Lesk, A. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advan. Enzymol.* **47**, 45–148.

Cohen, F. E. & Kuntz, I. D. (1987). Prediction of the three dimensional structure of human growth hormone. *Proteins: Struct. Funct. Genet.* **2**, 162–166.

Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, L. B., Ciardelli, T. L. & Smith, K. A. (1986). Structure-activity studies of interleukin-2. *Science*, **234**, 349–352.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., Mornon, J.-P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* **6**, 377–382.

Connolly, C. M. L. (1983*a*). Analytical molecular surface calculation. *J. Appl. Crystallogr.* **16**, 548–558.

Connolly, C. M. L. (1983*b*). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthase. *Proteins: Struct. Funct. Genet.* **2**, 118–129.

Curtis, B. M., Presnell, S. R., Srinivasan, S., Sassenfeld, H., Klinke, R., Jeffery, E., Cosman, D., March, C. J. & Cohen, F. E. (1991). Experimental and theoretical studies of the three-dimensional structure of human interleukin-4. *Proteins: Struct. Funct. Genet.* **11**, 111–119.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model for evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, p. 345, National Biomedical Research Foundation, Washington, DC.

Dever, T. E., Glynias, M. J. & Merrick, W. C. (1987). GTP-binding domain: three consensus sequence elements with distinct spacing. *Proc. Nat. Acad. Sci., U.S.A.*, **84**, 1814–1818.

de Vos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). Human growth hormone and extracellular domain of its receptor. *Science*, **255**, 306–312.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1981). Structure of bovine pancreatic phospholipase A2 at 1·7 Å resolution. *J. Mol. Biol.* **147**, 97–123.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature (London)*, **299**, 371–374.

Eisenberg, D., Wilcox, W., Eshita, S. M., Pryciak, P. M., Ho, S. P. & DeGrado, W. F. (1986). The design, synthesis, and crystallization of an alpha-helical peptide. *Proteins: Struct. Funct. Genet.* **1**, 16–22.

Eklund, H., Nordström, B., Zeppezauer, E., Söderlund, G., Ohlsson, I., Boiwe, T. , Söderberg, B. O., Tapia, O., Brändén, C. I. & Akeson, A. E. (1976). Three-dimensional structure of horse liver alcohol dehydrogenase at 2·4 Å resolution. *J. Mol. Biol.* **102**, 27–59.

Farber G. K. & Petsko G. (1990). The evolution of *α/β* barrel enzymes. *Trends Biochem. Sci.* **15**, 228–234.

Fasman, G. (1989). Editor of *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum, New York.

Fry, D. C., Kuby, S. A. & Mildvan, A. S. (1986). ATP-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, F_1-ATPase, and other nucleotide-binding proteins. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 907–911.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Gerloff, D. L., Jenny, T. F., Knecht, L. J., Gonnet, G. H. & Benner, S. A. (1993*a*). The nitrogenase MoFe protein: a secondary structure prediction. *FEBS Letters*, in the press.

Gerloff, D. L., Jenny, T. F., Knecht, L. J. & Benner, S. A. (1993*b*). A secondary structure prediction of the hemorrhagic metalloprotease family. *Biochem. Biophys. Res. Commun.* **194**, 560–565.

Gonnet, G. H. & Benner, S. A. (1991). *Computational Biochemistry Research at ETH.* Technical Report 154, Departement Informatik. 1-18.

Gonnet, G. H., Cohen, M. A. & Benner S. A. (1992*a*). Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.

Goraj, K., Renard, A. & Martial, J. A. (1990). Synthesis, purification and initial structural characterization of octarellin, a *de novo* peptide modelled on the α/β-barrel. *Protein Eng.* **3**, 259–266.

Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1·6 Å resolution. *J. Mol. Biol.* **169**, 521–563.

Gutte B., Daeumigen M. & Wittschieber E. (1979). Design, synthesis and characterization of a 34-residue polypeptide that interacts with nucleic acids. *Nature (London)*, **281**, 650–655.

Hahn, K. W., Klis, W. A. & Stewart, J. (1990). Design and synthesis of a peptide having chymotrypsin-like esterase activity. *Science*, **248**, 1544–1546.

Hecht, M. H., Richardson, J. S., Richardson, D. C. & Ogden, R. C. (1990). *De novo* design, expression, and characterization of felix: a four-helix bundle protein of native-like sequence. *Science*, **249**, 884–891.

Holbrook, S. R., Muskal, S. M. & Kim, S. H. (1990). Predicting surface exposure of amino acids from protein sequence. *Protein Eng.* **3**, 659–665.

Hopp, T. & Woods, K. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 3824–3828.

Hubbard, T. J. P. & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171.

Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988). Three-dimensional structure of the tryptophan synthase alpha$_2$beta$_2$ multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.* **263**, 17857–17871.

Johnsson, K., Allemann, R. K. & Benner, S. A. (1990). Designed enzymes: new peptides that fold in aqueous solution and catalyze reactions. In *Molecular Mechanisms in Bioorganic Processes* (Bleasdale, C. & Golding, B. T., eds), pp. 166–187, Royal Society of Chemistry, Cambridge.

Jörnvall, H., Persson, B. & Jefferey, J. (1987). Characteristics of alcohol/polyol dehydrogenases: the zinc-containing long-chain alcohol dehydrogenases. *Eur. J. Biochem.* **167**, 195–201.

Kaiser, E. T. (1988). In *Redesigning the Molecules of Life* (Benner, S. A., ed.), pp. 115–175, Springer-Verlag, Heidelberg.

Kaiser, E. T. & Kezdy, F. J. (1984). Amphiphilic secondary structure: design of peptide hormones. *Science*, **223**, 249–255.

Kim, J. & Rees, D. C. (1992). Crystallographic structure and functional implications of the nitrogenase molybdenum-iron protein from *Azotobacter vinelandii*. *Nature (London)*, **360**, 553–560.

Kimura, M. (1982). The neutral theory as a basis for understanding the mechanism of evolution and variation at the molecular level. In *Molecular Evolution, Protein Polymorphism, and the Neutral Theory* (Kimuara, M., ed.) pp. 3–56, Springer-Verlag, Berlin.

King, J. L. & Jukes, T. H. (1969). Non-Darwinian evolution: random fixation of selectively neutral mutations. *Science*, **164**, 788–798.

Knighton, D. R., Zheng, J., Ten Eyck, L., Ashford, F. V. A., Xuong, N. H., Taylor, S. S. & Sowadski, J. M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, **253**, 407–414.

Koyama, S., Yu, H., Dalgarno, D. C., Shin, T. B., Zydowsky, L. D. & Schreiber, S. L. (1993). Structure of PI3K SH3 domain and analysis of the SH3 family. *Cell*, **72**, 945–952.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Lenstra, J. A., Hofsteenge, J. & Beintema, J. J. (1977). Invariant features of the structure of pancreatic ribonuclease. *J. Mol. Biol.* **109**, 185–193.

Lesk, A. M. & Boswell, D. R. (1992). Does protein structure determine amino acid sequence? *BioEssays*, **14**, 407–410.

Lim, V. I. (1974*a*). Structural principles of the globular organization of protein chains: a stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88**, 857–872.

Lim, V. I. (1974*b*). Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* **88**, 873–894.

Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature (London)*, **339**, 31–36.

Maxfield, M. J. & Scheraga, H. A. (1979). Improvements in the prediction of protein backbone topography by reduction of statistical errors. *Biochemistry*, **18**, 697–704 (1979).

McKay, D. B. (1992). Response to unraveling the structure of IL-2. *Science*, **257**, 412–413.

Milburn, M. V., Prive, G. G., Milligan, D. L., Scott, W. G., Yeh, J., Jancarik, J., Koshland, D. E., Jr & Kim, S.-H. (1991). Three dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. *Science*, **254**, 1342–1346.

Moe, G. R. & Koshland, D. E., Jr (1986). Transmembrane signalling through the aspartate receptor. In *Current Communications in Molecular Biology: Microbial Energy Transduction: Genetics, Structure and Function of Membrane Proteins*, pp. 163–168, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Musacchio, A., Gibson, T., Lehto, V.-P. & Saraste, M. (1992*a*). SH3- an abundant protein domain in search of a function. *FEBS Letters*, **307**, 55–61.

Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992*b*). Crystal structure of a Src-homology 3 (SH3) domain. *Nature (London)*, **359**, 851–855.

Osterhout, J. J., Jr, Handel, T., Na, G., Toumadje, A., Long, R. C., Connolly, P. J., Hoch, J. C., Johnson, W. C., Jr, Live, D. & DeGrado, W. F. (1992). Characterization of the structural properties of α1B, a peptide designed to form a four-helix bundle. *J. Amer. Chem. Soc.* **114**, 331–337.

Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L.

(1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues, and structure prediction. *Proc. Roy. Soc. B.* **241**, 132–145.

Pace, N. R., Olsen, G. J. & Woese, C. R. (1986). Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell*, **45**, 325–326.

Padmanabham, S., Marqusee, S., Ridgeway, T., Laue, T. M. & Baldwin, R. L. (1990). Relative helix-forming tendencies of nonpolar amino acids. *Nature (London)*, **344**, 268–270.

Phillips, S. E. V. (1980). Structure and refinement of oxymyoglobin at 1·6 Å resolution. *J. Mol. Biol.*, **142**, 531–554.

Rees, D. C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science*, **245**, 510–513.

Robson, B. & Garnier, J. (1993). Protein structure prediction. *Nature (London)*, **361**, 506.

Rost, B. & Sander, C. (1992). Jury returns on structure prediction. *Nature (London)*, **360**, 540.

Rost, B., Schneider, R. & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem. Sci.* **18**, 120–123

Russell, R. B., Breed, J. & Barton, G. J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Letters*, **304**, 15–20 (1992).

Schiffer, M. & Edmundson, A. B. (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.* **7**, 121–135.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, pp. 1–314, Springer-Verlag, New York.

Sheehan, J. C., Bennett, G. B. & Schneider, J. A. (1966). Synthetic peptide models of enzyme active sites. III. Stereoselective esterase models. *J. Amer. Chem. Soc.* **88**, 3455–3456.

Shoji, S., Parmelee, D. C., Wade, R. D., Kumar, S., Ericsson, L. H., Walsh, K. A., Neurath, H., Long, G. L., Demaille, J. G., Fischer, E. H. & Titani, K. (1983). Complete amino acid sequence of the catalytic subunit of bovine cardiac muscle cyclic AMP-dependent protein kinase. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 848–851.

Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. *J. Mol. Biol.* **79**, 351–357.

Smith, D. L., Ringe, D., Finlayson, W. L. & Kirch, J. F. (1986). Preliminary X-ray data for aspartate aminotransferase from *Escherichia coli*. *J. Mol. Biol.* **191**, 301–302.

Sternberg, M. J. E. & Taylor, W. R. (1984). Modelling the ATP-binding site of oncogene products: the epidermal growth factor receptor and related proteins. *FEBS Letters*, **175**, 387–392.

Summers, N. L., Carlson, W. D. & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol.* **196**, 175–198.

Tainer, J. A., Getzoff, E. D., Beem, K. M., Richardson, J. S. & Richardson, D. C. (1982). Determination and analysis of the 2 Å structure of copper, zinc superoxide dismutase. *J. Mol. Biol.* **160**, 181–217.

Taylor, S. S., Buechler, J. A., Slice, L. W., Knighton, D. K., Durgerian, S., Ringheim, G. E., Neitzel, J. J., Yonemoto, W. M., Sowadski, J. M. & Dospmann, W. (1988). cAMP-dependent protein kinase: a framework for a diverse family of enzymes. *Cold Spring Harbor Symp. Quant. Biol.* **53**, 121–130.

Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1991). Prediction of progress at last. *Nature (London)*, **354**, 105–106.

Waksman, G., Kominos, D., Robertson, S. C., Pant, N., Baltimore, D., Birge, R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., Resh, M. D., Rios, C. B., Silverman, L. & Kuriyan, J. (1992). Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine phosphorylated peptides. *Nature (London)*, **358**, 646–653.

Weinhold, E. G., Glasfeld, A., Ellington, A. D. & Benner, S. A. (1991). Structural determinants of stereospecificity in yeast alcohol dehydrogenase. *Proc. Nat. Acad. Sci. U.S.A.* **88**, 8420–8424.

White, J. L., Hackert, M. L., Buehner, M., Adams, M. J., Ford, G. C., Lentz, P. J., Jr, Smiley, I. E., Steindel, S. J. & Rossmann, M. G. (1976). A comparison of the structures of apo dogfish M4 lactate dehydrogenase and its ternary complexes. *J. Mol. Biol.* **102**, 759–779.

Yu, H., Rosen, M. K., Shin, T. B., Seidel-Dugan, C., Brugge, J. S. & Schreiber, S. L. (1992). Solution structure of the SH3 domain of Src and identification of its ligand-binding site. *Science*, **258**, 1665 (1992).

Zvelebil, M. J. & Sternberg, M.J.E. (1988). Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.* **2**, 127–138.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.

Edited by A. R. Fersht

(Received 23 October 1992; accepted 10 August 1993)