doi:10.1006/jmbi.2001.4540 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **307**, 1487–1502

JMB

Three-dimensional Cluster Analysis Identifies Interfaces and Functional Residue Clusters in Proteins

Ralf Landgraf, Ioannis Xenarios and David Eisenberg*

UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, 405 Hilgard Avenue, Box 951570 Los Angeles, CA 90095-1570 USA

Three-dimensional cluster analysis offers a method for the prediction of functional residue clusters in proteins. This method requires a representative structure and a multiple sequence alignment as input data. Individual residues are represented in terms of regional alignments that reflect both their structural environment and their evolutionary variation, as defined by the alignment of homologous sequences. From the overall (global) and the residue-specific (regional) alignments, we calculate the global and regional similarity matrices, containing scores for all pairwise sequence comparisons in the respective alignments. Comparing the matrices yields two scores for each residue. The regional conservation score ($C_R(x)$) defines the conservation of each residue x and its neighbors in 3D space relative to the protein as a whole. The similarity deviation score ($S(x)$) detects residue clusters with sequence similarities that deviate from the similarities suggested by the full-length sequences. We evaluated 3D cluster analysis on a set of 35 families of proteins with available cocrystal structures, showing small ligand interfaces, nucleic acid interfaces and two types of protein-protein interfaces (transient and stable). We present two examples in detail: fructose-1,6-bisphosphate aldolase and the mitogen-activated protein kinase ERK2. We found that the regional conservation score ($C_R(x)$) identifies functional residue clusters better than a scoring scheme that does not take 3D information into account. $C_R(x)$ is particularly useful for the prediction of poorly conserved, transient protein-protein interfaces. Many of the proteins studied contained residue clusters with elevated similarity deviation scores. These residue clusters correlate with specificity-conferring regions: 3D cluster analysis therefore represents an easily applied method for the prediction of functionally relevant spatial clusters of residues in proteins.

Keywords: bioinformatics, evolutionary tracing, protein families, residue patches, phylogeny

**Corresponding author*

Introduction

The prediction of functionally relevant residues in proteins can lead to the assignment of new functions and elucidate the mechanism by which proteins carry out known functions. Various methods have been applied to this important problem. Some of the predictions are based on biophysical properties of individual residues (Jones & Thornton, 1997a,b; Tsai *et al*., 1997; Xu *et al*., 1997), others are based on harvesting evolutionary information inherent in sets of homologous sequences. Of those methods harvesting the evolutionary information, the most direct approach (Bucher & Bairoch, 1994; Henikoff & Henikoff, 1991) assigns sequence motifs directly to particular functions. In a more elaborate approach (Casari *et al*., 1995), vectorial analysis of sequence profiles is used to identify functionally important residues. In a third method, evolutionary tracing (Lichtarge *et al*., 1996, 1997), information inherent in a phylogenetic tree is added to the analysis of conserved sequences, often revealing more subtle aspects of protein function. Starting with a multiple sequence alignment,

Abbreviations used: ERK, extracellular signal-regulated kinase; MAPK, mitogen-activated protein kinase; FBP, fructose 1,6-bisphosphate; F-1-P, fructose 1-phosphate; DHAP, dihydroxyacetone phosphate; $C_R(x)$, regional conservation score for residue x; $S(x)$, similarity deviation score; $C_P(x)$, positional conservation score for residue x; CD, common docking.

E-mail address of the corresponding author: david@mbi.ucla.edu

0022-2836/01/051487–16 $35.00/0

a representative structure, and a phylogenetic tree, this method evaluates the conservation at each position in the alignment for different sequence similarity cut-offs. In its original implementation, residues are classified as variable, conserved or group-specific, that is specific to one branch of the phylogenetic tree. This analysis can be further expanded by the use of amino acid substitution matrices to evaluate conservation (Landgraf *et al*., 1999). In either case, a representative structure is used to visualize the distribution of scores at the end of the analysis.

Here, we present 3D cluster analysis, a further extension of evolutionary tracing (Lichtarge *et al*., 1997) with two goals. The first goal is to improve the sensitivity with which functional residue clusters can be identified. The availability of a representative structure provides a source of significant additional information, which can improve the sensitivity of detection. In addition to projecting the final scores onto a reference structure, 3D cluster analysis makes the structural information an integral part of the analysis. The second goal is to identify functionally relevant residue clusters within a protein without reliance on a phylogenetic tree as input data. The grouping of sequences in a phylogenetic tree often reflects similarity in function, a fact that is exploited in evolutionary tracing. However, we speculate that a protein possesses regions or residue clusters for which the phylogenetic tree does not adequately reflect relationships of sequence similarity and function. We foresee several scenarios in which such a deviation might occur. For a protein with multiple functional residue clusters, the grouping of sequence in the apparent phylogenetic tree can represent the average of several conserved functions. In addition, the similarity relationships of a highly conserved residue cluster could dominate the phylogenetic tree and overshadow the grouping suggested by a less-conserved residue cluster associated with a different function. The detection of such clusters, associated with secondary functions of the protein, would not be possible by conventional evolutionary tracing. Here, we propose a score, termed similarity deviation score ($S(x)$), which detects residue clusters that exhibit deviations in their regional sequence similarity relationships. The detection of such residue clusters should facilitate the assignment of functions that are not adequately represented in the grouping of the ''apparent phylogenetic tree''.

Three-dimensional cluster analysis places structural information at the core of the analysis and evaluates conservation in terms of spatially defined residue clusters within a protein. This method requires a representative structure and multiple sequence alignment but no phylogenetic tree as input. Our analysis shows that functionally relevant residue clusters that exhibit a low degree of conservation can be detected with enhanced sensitivity when we use regional conservation scores as opposed to a scoring scheme that does not take 3D information into account. We also find cases of residue clusters that, when compared between different sequences, show similarity relationships that deviate from the similarity relationships observed for the protein as a whole. Comparison with biochemical data suggests that these residue clusters confer specificity to catalytic reactions or protein interactions.

We evaluate 3D cluster analysis on a set of 35 protein families, for which a cocrystal structure of a representative member identifies functionally relevant interfaces. We first analyze what percentage of known interface residues can be identified by 3D cluster analysis. We then evaluate the predictions made by 3D cluster analysis against the known biochemical properties of two families of proteins in detail. The first family is extracellular signal-regulated kinase (ERK) 2. The ERK1/ERK2 mitogen-activated protein (MAP) kinases (MAPKs) represent one of four known MAP kinase pathways in mammalian cells, where they transduce signals in response to various growth factors (Blenis, 1993; Blumer & Johnson, 1994; Davis, 1993; Schlessinger, 1994). ERK2 is activated through phosphorylation by upstream kinases (MAPKKs) such as MEK1/MEK2 (Crews *et al*., 1992; Mansour *et al*., 1994) and is subject to deactivation by phosphatases (Anderson *et al*., 1990; Boulton & Cobb, 1991; Zheng & Guan, 1993). MAPKKs show strong specificity in their interaction with MAPKs and the nature of this specificity has been an area of intense investigation.

The second family of proteins to which we apply 3D cluster analysis are type I fructose-1,6-bisphosphate aldolases (aldolases). Type I aldolases catalyze the Schiff base-mediated, reversible cleavage of fructose 1,6,-bisphosphate (FBP) or fructose 1-phosphate (F-1-P) to dihydroxyacetone phosphate (DHAP) and glyceraldehyde 3-phosphate or glyceraldehyde, respectively (for a general review, see Horecker *et al*., 1972). The functional unit of aldolase is a tetramer and the active sites in each monomer are found at the core of the monomer $(\beta/\alpha)_8$ barrel (Cooper *et al*., 1996; Dalby *et al*., 1999; Gamblin *et al*., 1990; Sygusch *et al*., 1987). Mammalian class I aldolases exist in three isoforms with different tissue distribution (type A, muscle; B, liver; C, brain) and catalytic preference for FBP over F-1-P. The liver (B) isoform stands out, in that it utilizes both substrates equally well (Penhoet *et al*., 1966, 1969), a reflection of the utilization of exogenous fructose by the liver enzyme. In addition to its catalytic activity, aldolase has been reported to interact with the cytoskeleton. The metabolically controlled interaction of aldolase with F and G actin has been shown to mediate the association of the insulin-responsive glucose transporter (Glut-4) with the cytoskeleton, thereby regulating the levels of glucose transporter molecules on the cell surface (Clarke & Masters, 1975; Clarke & Morton, 1976, 1982; Clarke *et al*., 1984; Walsh *et al*., 1981).

Results

For a protein family, 3D cluster analysis evaluates the residue conservation at each position of a reference structure. The conservation is evaluated for the 3D region surrounding residue x as defined by the spatial neighbors of x in the reference structure. The regional conservation is calculated and assigned to residue x. Also, the similarity deviation score is calculated. This is the extent to which the sequence similarity relationships for this residue cluster deviate from that of the protein as a whole. Three-dimensional cluster analysis includes the following basic steps, as outlined in Figure 1.

(1) Selection of a reference structure for the family of proteins under investigation and identification of sequences with high sequence similarity (judged by *E*-score) to the reference protein.

(2) Creation of a (global) multiple sequence alignment based on the full-length sequence of the reference protein.

(3) Identification of structural neighbors for each residue x in the reference structure with C^{α} atoms within a set radius (default 10 Å, Step I in Figure 1). An option is to evaluate only surface-exposed residues as neighbors.

(4) Extraction of regional alignments from the global alignment. One regional alignment for each residue in the reference structure, containing its structural neighbors (Step II in Figure 1). The length of the alignment is equal to the number of neighbors for this residue.

(5) Calculation of global and regional similarity matrices, containing the pairwise sequence similarity scores for all sequences within the respective alignments (Step III in Figure 1).

(6) Calculation of the regional conservation score ($C_R(x)$) for each residue x in the reference structure, representing the difference in conservation between the structural neighbors of residue x *versus* the protein as a whole.

(7) Calculation of the similarity deviation score ($S(x)$) reflecting the degree of correlation between the regional and global similarity matrix. A high $S(x)$ score indicates a strong deviation between the similarity relationships within the regional

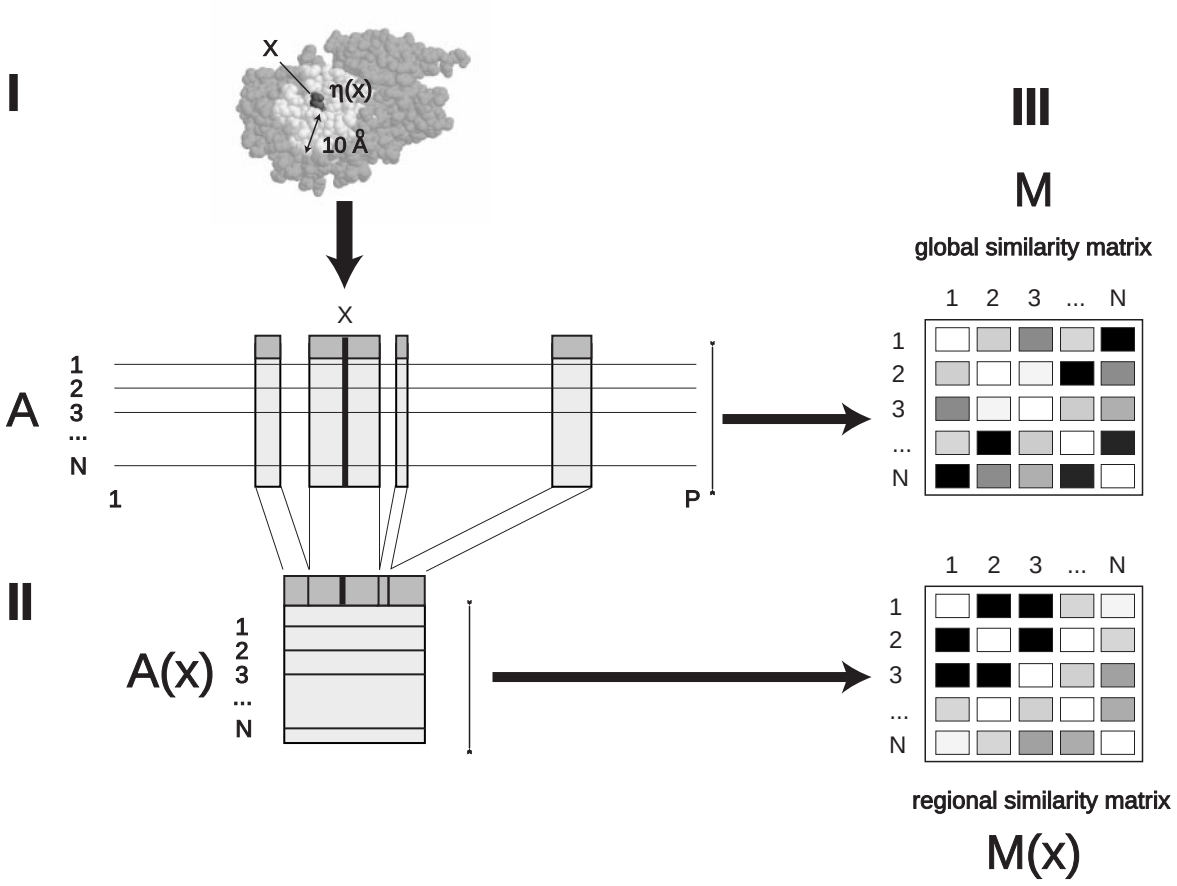

Figure 1. Basic steps in 3D cluster analysis. The extraction of regional alignments for each residue in the reference structure links structural information to the sequence alignment. (I) For each residue x, all structurally adjacent residues within a given radius (e.g. 10 Å) are identified. (II) The identified positions (highlighted as gray blocks) are extracted from the global alignment A. These blocks are joined to form a regional alignment with N sequences. (III) Two similarity matrices of dimension $N \times N$ are generated, a global similarity matrix (M) representing the relationship of all full-length sequences and a regional similarity matrix ($M(x)$) representing the relationship of all sequences in the regional alignment, $A(x)$.

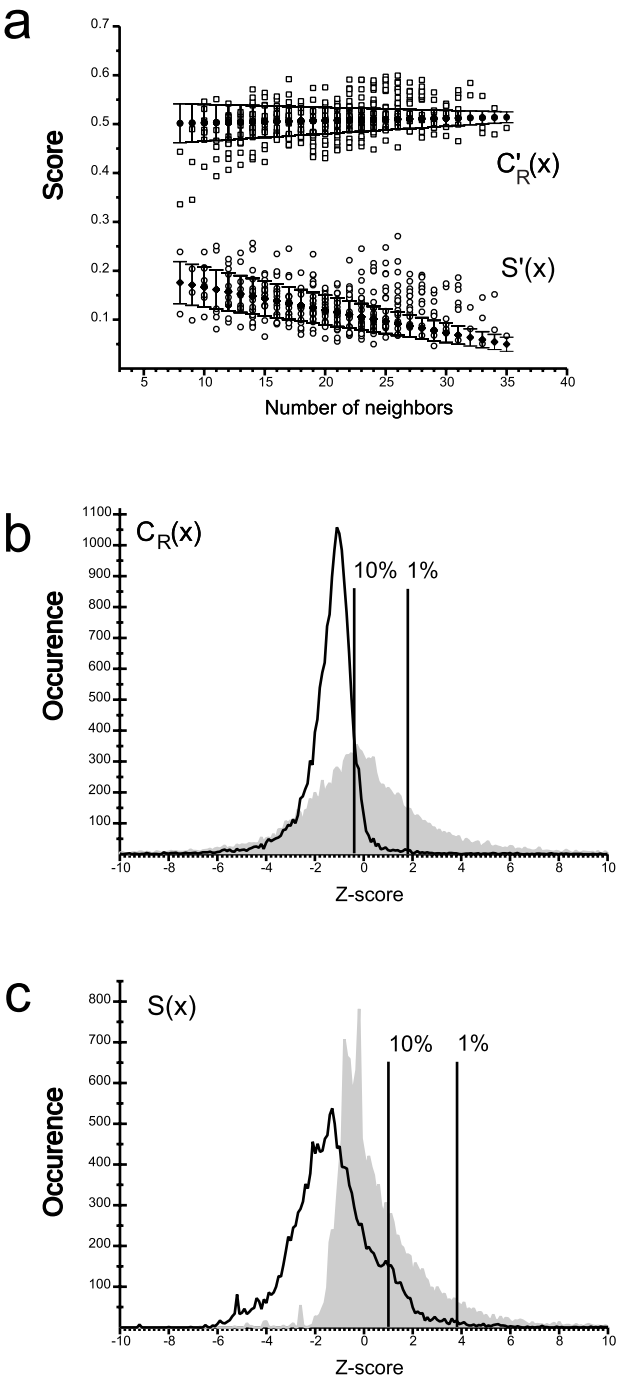

Figure 2. The distributions of $C_R(x)$ and $S(x)$ scores are distinct from the distributions of scores based on reshuffled alignments. (a) Distribution of the raw $C'_R(x)$ (□) and $S'(x)$ scores (○) and the respective scores derived from randomly assembled neighborhoods (mean of random $S'(x)$ (◆), $C'_R(x)$ (●). Standard deviation of random neighborhood scores is indicated as error bars). The raw $C'_R(x)$ and $S'(x)$ scores were converted to Z-scores, using the mean and standard deviation for a random neighborhood containing an equal number of residues. The raw data shown were obtained for aldolase. (b) and (c) Distribution of regional conservation Z-scores ($C'_R(x)$, (b) and similarity deviation Z-scores ($S(x)$, (c) (gray area graph) compared to the distribution of the respective scores obtained from reshuffled alignments ($rC_R(x)$ and $rS(x)$, continuous line). The distribution of Z-scores from all residues in all 35 protein families is shown. The Z-scores (background thresholds) at which the distribution of scores from reshuffled alignments has less than 1 or 10% of residues above the threshold is indicated in both cases.

alignment of the structural neighbors of residue x and the similarity relationships obtained for the full-length sequences.

(8) Visualization of the distribution of regional conservation and similarity deviation scores based on the reference structure.

Determining background thresholds for $C_R(x)$ and $S(x)$ Z-scores

We applied 3D cluster analysis to 35 protein families. The protein families selected represent four classes of proteins: proteins with small ligand-binding sites, RNA-binding proteins, DNA-binding proteins and protein complexes. Protein complexes were further classified as stable (e.g. the aldolase tetramer) or transient (e.g. the Ras-Gap complex). Figure 2(a) shows the raw $C'_R(x)$ and $S'(x)$ scores for the aldolase protein family as a function of the number of neighboring residues within a 10 Å radius of each evaluated residue. We compare the raw scores to the mean and standard deviation (error bars in Figure 2(a)) of control scores obtained for an equal number of randomly selected, non-neighboring residues within the reference structure. The raw $C_R'(x)$ scores cluster around 0.5, the expected value for comparable conservation within the regional and local similarity matrix. For the raw $S'(x)$ score, values gradually approach zero with an increasing number of evaluated residues, because the difference between the regional and the global similarity matrix diminishes. However, as is the case for the $C'_R(x)$ scores, the data obtained for ''true'' structure-based neighbors show a broader distribution than the standard deviation of scores obtained for randomly selected ''neighbors''.

Based on the comparison of the raw $C'_R(x)$ and $S'(x)$ scores and the distribution of scores from randomly assembled residue ''neighborhoods'', we calculated Z-scores ($C_R(x)$) and $S(x)$). Figure 2(b) and (c) show the distribution of Z-scores for all protein families evaluated. To determine the threshold above which a Z-score can be considered to indicate a true regional conservation or regional deviation of sequence similarity, we generated a second set of Z-scores based on randomly reshuffled alignments ($rS(x)$ and $rC_R(x)$). For both, the $S(x)$ and $C_R(x)$ scores, the distribution of scores obtained from the reshuffled alignments are distinct, although a substantially larger separation exists in the case of the $C_R(x)$ score (Figure 2(b)). We took the percentage of $rS(x)$ and $rC_R(x)$ scores above a specific cut-off to be an indicator of the

number of anticipated false positives, and extracted the *Z*-score above which less than 1% and 10%, respectively, of scores can be observed for r$S(x)$ and r$C_R(x)$. We refer to these *Z*-scores as the 1% and 10% background thresholds. For the r$C_R(x)$ score, the background thresholds were 1.8 and −0.4, respectively (Figure 2(b)). At these *Z*-scores, 18% and 51%, respectively, of the $C_R(x)$ scores are above the threshold. In the case of the r$S(x)$ score (Figure 2(c)), the 1% and 10% background thresholds were 3.8 and 1.0 with 7% and 31%, respectively, of the $S(x)$ scores above the threshold. This difference in the extent to which the $C_R(x)$ and $S(x)$ vary from the r$S(x)$ and r$C_R(x)$ scores is also apparent when raw scores (equivalent to the aldolase data in Figure 2(a)) are plotted separately for each protein family (data not shown). The distribution seen for the $C'_R(x)$ scores of aldolase in Figure 2(a) are representative of most protein families analyzed. However, the extent to which the raw $S(x)$ scores show a broader distribution than the standard deviation of scores for random neighbors differs significantly between protein families.

Testing the ability of 3D cluster analysis to detect interfaces

Next we evaluated the ability of the 3D cluster analysis to identify known clusters of functional residues. For this, we tabulated the extent to which residue positions known to be in interfaces receive *Z*-scores above the background thresholds determined above. Residues involved in interfaces with small ligands, nucleic acids or other proteins were identified from cocrystal structures. This class of residues represents only a subset of functionally relevant residues and is limited to those residues making direct ligand contact, but has the advantage of a clear standard for evaluation, namely the percentage of buried surface area upon complex formation. Table 1 lists the number of residues in the various types of interfaces and the extent to which these residues were identified by 3D cluster analysis. To evaluate the extent to which the incorporation of 3D information benefits the prediction of functional residues, we contrast the results obtained for the regional conservation score (3D information included) with a positional conservation score ($C_P(x)$, no 3D information included), described in Materials and Methods. In brief, this score measures the conservation at each position in an alignment without consideration of neighboring residues in the structure. No randomization scheme equivalent to that used for the generation of the $C_R(x)$ *Z*-score is available for the $C_P(x)$ score. In order to find a comparable background threshold for the $C_P(x)$ score, we created a histogram of $C_P(x)$ scores for each protein family and determined the score above which the percentage of residues with scores above the thresholds is equivalent to the percentage of residues above the $C_R(x)$ background threshold. The absolute value of the background threshold for the $C_p(x)$ score will therefore differ for each example, depending on the degree of overall sequence conservation within each protein family.

Table 1 compares for four types of interfaces and two *E*-score thresholds the percentage of known interface residues above background thresholds for the three scores ($C_R(x)$, $S(x)$ and $C_p(x)$). The results are compared with those obtained on the basis of reshuffled alignments (shown in parentheses) to obtain a measure for the percentage of anticipated false positives. Several features emerge from this comparison. The $C_R(x)$ scores identify, on average, 36% of interface residues at the most stringent background threshold (<1% expected from reshuffled alignments) and 67% at a less stringent background threshold, anticipating less than 10% of the high scores to occur at random. The identification of interface residues increases as the sequence diversity increases (*E*-score 10^{-50} *versus* 10^{-20}), in

Table 1. The 3D cluster analysis identifies the majority of residues in interfaces

	Catalytic sites (small molecule complexes) 15 examples 214 residues		Protein-DNA/RNA complexes 6 examples 68 residues		Protein-protein complexes: Stable 12 examples 381 residues		Protein-protein complexes: Transient 13 examples 208 residues		
Score	1%	10%	1%	10%	1%	10%	1%	10%	*E*-score
$C_R(x)$	50 (1)	83 (2)	31 (0)	62 (4)	27 (1)	64 (6)	17 (0)	53 (0)	10^{-50}
$C_P(x)$	66 (19)	90 (55)	62 (16)	77 (47)	36 (18)	58 (51)	46 (15)	61 (50)	10^{-50}
$S(x)$	27 (0)	62 (9)	11 (0)	22 (7)	5 (1)	30 (15)	15 (2)	52 (4)	10^{-50}
$C_R(x)$	63 (0)	88 (19)	39 (0)	66 (19)	28 (1)	63 (10)	31 (1)	72 (6)	10^{-20}
$C_P(x)$	54 (20)	86 (50)	50 (21)	66 (60)	32 (20)	59 (54)	16 (17)	51 (45)	10^{-20}
$S(x)$	32 (0)	70 (18)	2 (0)	22 (4)	7 (1)	25 (19)	7 (0)	38 (22)	10^{-20}

The four categories of interfaces are listed together with the number of example proteins and the total number of interface residues in this category. For each category we list the percentage of interface residues with scores above the two background thresholds (1 or 10%) indicated in Figure 2(b) and (c). For comparison, the percentage of residues with scores above the background threshold, calculated for this particular interface and based on reshuffled alignments, is given in parentheses. Results are presented for two different *E*-score thresholds.

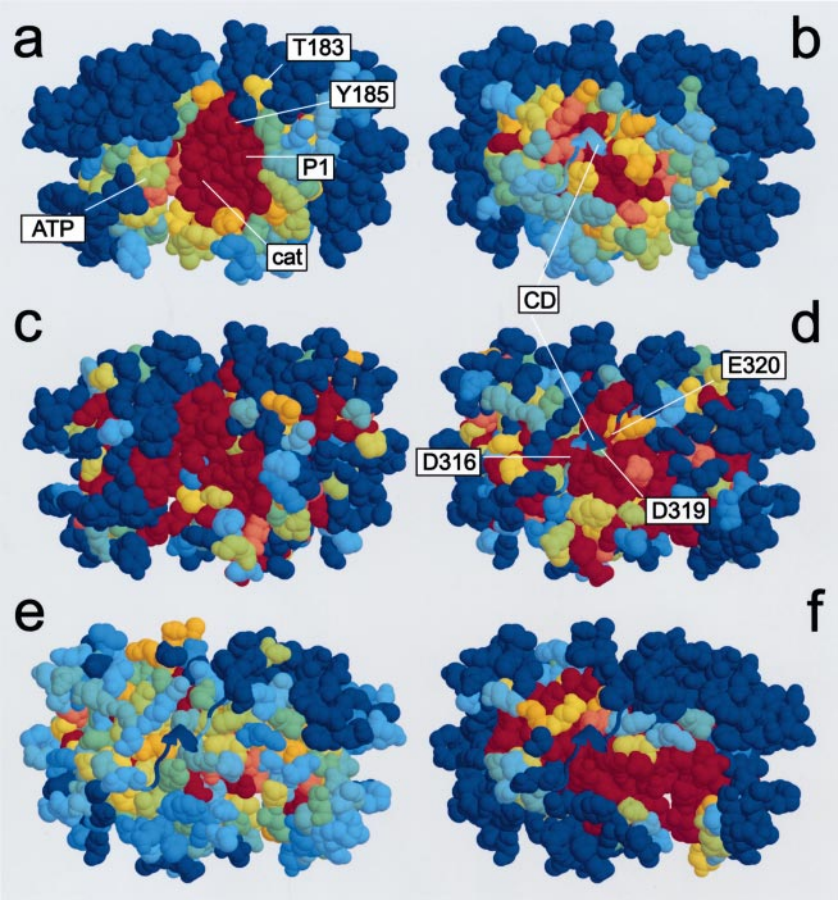

Figure 3. The regional conservation score identifies two distinct interfaces on ERK2 and benefits from the inclusion of more distant homologs. The regional conservation scores ($C_R(x)$) and positional conservation scores ($C_P(x)$) ((c) and (d)) are superimposed on the structure of rat ERK2. Scores above the 1% *Z*-score threshold (or $C_p(x)$ score equivalent) are presented in red. All scores below the 10% background threshold are colored in dark blue. All remaining residues are colored on a spectral scale from orange to light blue with decreasing scores. (a) and (c) Show the ''front'' of ERK2, containing the key catalytic region (cat), the P1 site (P1), ATP-binding pocket (ATP) and the dual phosphorylation site (T183, Y185). The key feature on the backside of the protein ((b) and (d)) is the L16 loop (shown in ribbon representation). Key residues of the common docking domain are indicated. The regional conservation score identifies a conserved residue cluster on the front of ERK2 and the improved signal to noise ratio allows one to define the outlines of a conserved residue cluster on the backside of ERK2 more clearly. (e) and (f) Backside of ERK2 with superimposed (e) $C_p(x)$ and (f) $C_R(x)$ (f) scores, calculated with an *E*-score threshold of 10^{-20}. In contrast to the positional conservation score, the inclusion of more distant homologs improves the identification of the backside interface on ERK2 by the $C_R(x)$ score.

some categories at the expense of accuracy, especially at the less-stringent background threshold. The interface categories show marked differences with regard to the percentage of recovered residues and the percentage of false positives. A high recovery with low error rates can be achieved using the 10% background threshold and a stringent *E*-score threshold (10^{-50}) in the case of catalytic sites, protein-nucleic acid interfaces, and even stable protein-protein interfaces. Notably, the prediction of transient protein-protein interfaces benefits the most from the inclusion of more distant sequence homologs. At a background threshold, allowing for up to 10% false positives and an *E*-score threshold of 10^{-20}, 72% of the transient protein interface residues can be recovered with a background of only 6% from reshuffled alignments.

Comparison of the predictive value of the different scores

A comparison of the $C_R(x)$ and $C_p(x)$ score shows that the $C_p(x)$ score identifies a large number of the interface residues, especially in the catalytic sites. However, the information value of this score for the prediction of functional residue clusters is diminished by a high rate of false positives. A comparison with the scores obtained from reshuffled alignments suggests that the $C_p(x)$ score can be used successfully for active sites and highly conserved residues in protein-DNA interfaces but is unlikely to provide good predictions in the case of less conserved protein-protein interfaces. In contrast to the $C_R(x)$ score, the addition of more distant sequence homologs diminishes the predictive power of the $C_p(x)$ score. In short, the $C_R(x)$ score is generally more effective than the $C_p(x)$ score in finding clusters of functional residues.

At the outset of this analysis, we asked whether regional differences in sequence similarity relationships exist. The tabulated results for the $S(x)$ score show that a significantly higher percentage of interface residues has $S(x)$ scores above the background threshold compared to the percentage obtained from reshuffled alignments. As is the case for the $C_R(x)$ score, marked differences exist between the four categories. In the case of the $S(x)$ score, we observe the largest percentage for active sites and transient interfaces. The percentages obtained for protein-nucleic acid and stable

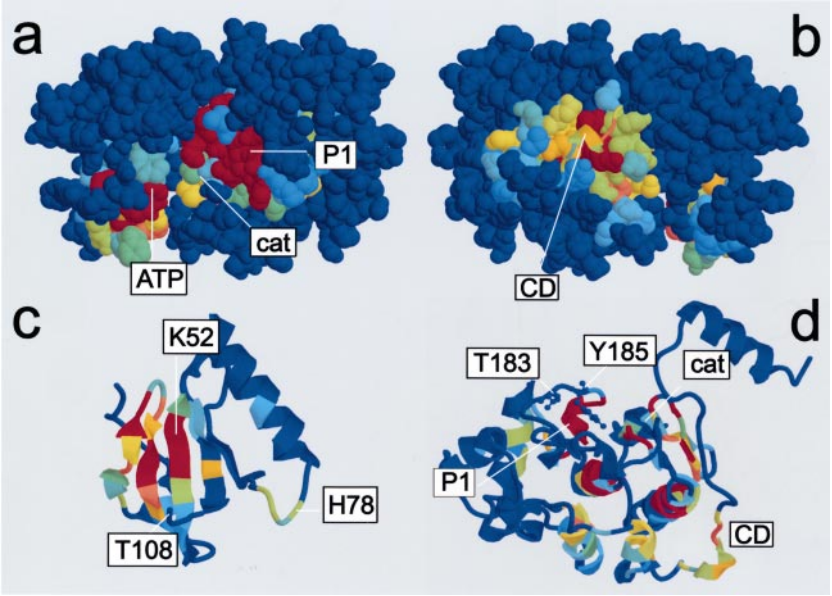

Figure 4. The similarity deviation score identifies three distinct residue clusters in ERK2. (a) On the front, the specificity conferring P1 site and the ATP binding pocket receive high scores. (b) On the backside, the $S(x)$ scores outlines a narrow cluster of residues, centered on the common docking (CD) domain. (c) Ribbon diagram of the N-terminal domain of ERK2 with superimposed similarity deviation scores and key residues within the ATP-binding site. (d) Top view of the C-terminal domain with superimposed similarity deviation scores. A cluster of high scores is located in the center of a bundle of helices forming the core and backside of ERK2. (The coloring scheme is as described for Figure 3).

protein-protein interfaces are significantly lower. In the case of catalytic sites and transient interfaces, the best ratio of identified residues compared to anticipated false positives is obtained for the more stringent *E*-score threshold (10^{-50}). However, the $S(x)$ score may highlight functional regions of proteins other than interfaces. To evaluate the prediction of other functional residue clusters requires further study of the biochemistry for each protein family. We selected two protein families, to investigate the predictive power of 3D cluster analysis in more detail.

The 3D cluster analysis of the MAP kinase ERK2

The reference structure for this analysis is that of rat ERK2 (Gamblin *et al.*, 1990). A FASTA search with the sequence of rat ERK2 identified 147 sequences with an *E*-score below 10^{-50}. The 147 homologous sequences are primarily composed of close ERK homologs, homologs of the related P38 MAPK and a series of MAPKs from higher plants (Figure 5). Figure 3 shows the regional conservation ($C_R(x)$, Figure 3(a) and (b)) and positional conservation score ($C_P(x)$, Figure 3(c) and (d)). Conserved features on the front of the larger C-terminal lobe (right side of Figure 3(a) and (c)), such as the phosphorylation lip with Thr183 and Tyr185, the catalytic region (residues 147 to 152), and the substrate specificity-conferring P1 site (residues 186 to 192), are readily detectable in both scoring schemes. However, residues with high positional conservation scores are more scattered, especially on the ''backside''(Figure 3(d)). In contrast, the regional conservation score (Figure 3(b)) clearly identifies the outlines of a conserved residue cluster, centered on the poorly conserved L16 loop (shown in ribbon representation). This L16 loop contains the recently identified common docking (CD) domain (Tanoue *et al.*, 2000). At an *E*-score threshold of 10^{-20}, the regional conservation score defines the outlines of a contiguous residue cluster even more clearly (Figure 3(f)). In contrast, the signal on the backside is all but lost when we use the positional conservation score (Figure 3(e)) at this *E*-score threshold. This observation confirms our earlier finding that the $C_R(x)$ score, in contrast to the $C_P(x)$ score, benefits from the inclusion of distant homologs.

Due to the inclusion of neighboring residues into the regional alignments, ''signal spill-over'' could occur from an adjacent residue cluster with high scores. To confirm that the elevated conservation scores on the backside of ERK2 is not the result of signal spill-over from high-scoring residues on the front of ERK2, we included only neighbors with a minimum surface exposure of 3 Å^2 in the analysis. This approach removes residues in the core of ERK2 that could facilitate signal spill over from the front of the molecule. Compared to the calculation without a limit for minimum surface exposure, only small changes could be observed on the backside of ERK2, confirming that the identified residue cluster is indeed an independent residue cluster.

A comparison of the regional conservation scores (Figure 3) and similarity deviation scores (Figure 4) shows considerable overlap with respect to the broadly defined residue clusters but differences in the emphasis of subsections. On the front of ERK2 (Figure 4(a)), the catalytic region and phosphorylation lip have been de-emphasized by the $S(x)$ score, while the highest scores can be seen for the specificity-conferring P1 region and the ATP-binding pocket. The P1 site confers specificity for substrates with proline adjacent to the phosphorylation site (position P + 1). On the backside, a narrowly defined cluster of residues shows high $S(x)$ scores (Figure 4(b)). This residue cluster consists of

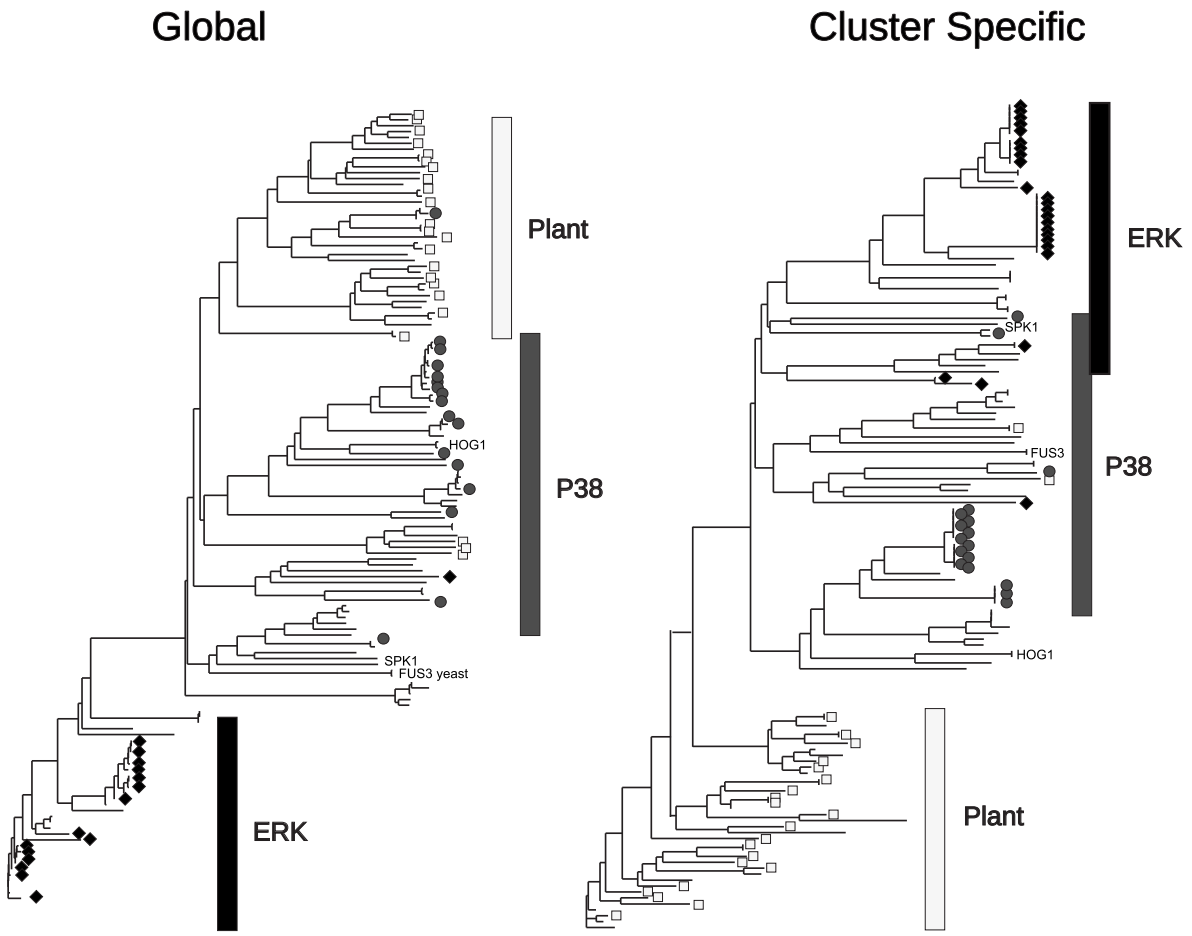

Figure 5. The N-terminal ATP-binding site of ERK2 shows shifted sequence similarity relationships. The shift in sequence similarity relationships is evident in the apparent phylograms derived from the full-length sequence (global) and N-terminal residues with high $S(x)$ scores (cluster specific). Homologs of ERK2 are marked as black diamonds, P38 homologs dark gray circles, and plant MAPKs as light gray squares. The MAPKs HOG1, SPK1 and FUS3 are labeled individually.

the CD domain containing L16 loop and its immediate surrounding.

The N-terminal domain of ERK2 primarily scores based on its $S(x)$ score. Further analysis of the N-terminal domain of ERK2, shown in ribbon representation in Figure 4(c), identifies the N-terminal β-sheet as the source of elevated $S(x)$ scores. This β-sheet matches the PROSITE profile for ATP binding sites in protein kinases (PS00107). Mutational analysis implicates Lys52 in ATP-binding (Robinson *et al.*, 1996). We extracted all high-scoring residues from this residue cluster (residues 23 to 41, 49 to 55 and 102) and compared the ''global phylogram'', based on the full-length sequence, with a ''cluster-specific phylogram'', based on the comparison of residues from this residue cluster alone (Figure 5). The global phylogram clusters all ERK homologs, while homologs of the mammalian P38 MAPK, and plant MAPKs are clearly set apart. In contrast, the cluster-specific phylogram groups ERK and P38 homologs together and sets plant MAPKs aside. The ATP binding of ERK2 site is therefore one example of a functionally relevant residue cluster that exhibits sequence similarity relationships that deviate form those derived from the full-length sequence.

The 3D cluster analysis of aldolase

As a second example for a detailed evaluation of 3D cluster analysis predictions, we chose type I fructose-bisphosphate aldolase (aldolase). The reference structure for this analysis was that of one subunit of the rabbit muscle aldolase tetramer (Blom & Sygusch, 1997). A FASTA search with an *E*-score threshold of 10^{-50} provided 109 homologous sequences, representing all three mammalian isoforms as well as plant, *Drosophila* and several parasite aldolase sequences. The results of the 3D cluster analysis of aldolase are summarized in Figure 6. Compared to ERK2, positions with high positional conservation scores are even more scattered throughout the structure (Figure 6(a)). Although key residues show high $C_p(x)$ scores, the

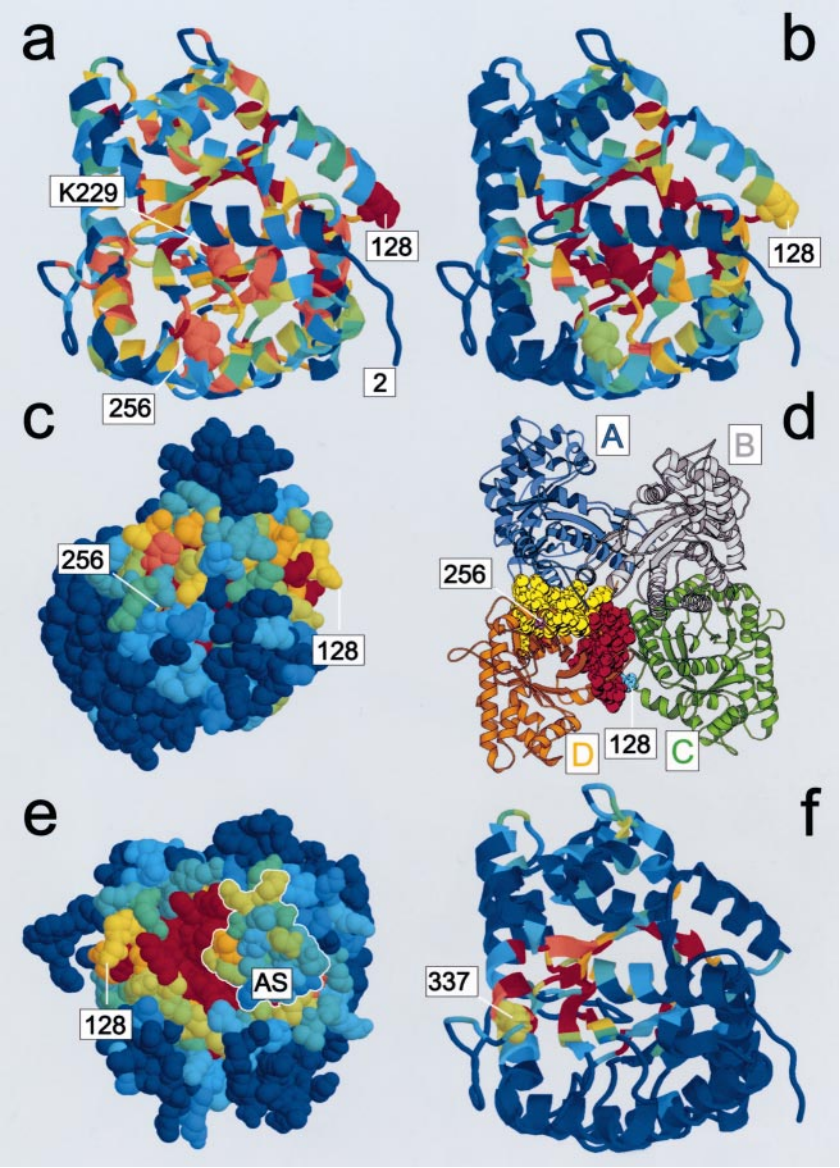

Figure 6. The 3D cluster analysis of aldolase identifies oligomer interfaces and a specificity-conferring residue cluster. (a) and (b) The use of regional conservation scores improves the signal to noise ratio compared to positional conservation scores. Scores are superimposed on the structure of the monomer of rabbit muscle aldolase (the coloring scheme is as described for Figure 3). (a) Positional conservation scoring results in scattered high scores. (b) Regional conservation scores identify the catalytic core region (Schiff base forming Lys229 in CPK) and marginally conserved interfaces. The identified oligomer interfaces are marked by positions 128 and 256 (shown as CPK). Mutations in these positions are known to disrupt tetramer formation. (c) The CPK representation of the regional conservation scores emphasizes the clear outlines of the distinct oligomer interfaces. The orientation of aldolase is similar to that of subunit D in the aldolase tetramer, but is slightly rotated to present a better frontal view of both interfaces. (d) Structure of the aldolase tetramer with highlighted interfaces (CPK), as defined by the regional conservation scores, and location of interface-disrupting mutations in position 256 (magenta) and 128 (cyan). (e) CPK presentation of aldolase monomer with superimposed $C_R(x)$ scores showing the backside view relative to (c). In addition to portions of the A:B (C:D) interface around residue 128 and the highly conserved entrance to the active site (center), a marginally conserved region to the right contains the residues implicated in the interaction of aldolase with actin (AS, outlined in white). (f) The similarity deviation scores identify portions of the active site and a residue cluster near the C terminus. Mutations in positions 337 (CPK) are associated with a loss of liver isoform specificity.

scatter of the signal makes it difficult to identify additional functional residue clusters.

A significant improvement in the signal to noise ratio can be achieved when $C_R(x)$ scores are evaluated instead of $C_P(x)$ scores. The area showing the highest $C_R(x)$ scores is located in the core of the β/α-barrel and includes all key residues known to be involved in catalysis (Figure 6(b), Schiff base forming Lys229 in CPK representation). The reduced scattering of signal facilitates the identification of contiguous residue clusters with low to intermediate conservation scores. Two such residue clusters are found at the periphery. Both clusters are associated with naturally occurring mutations in positions 256 and 128, known to interfere with the association of aldolase monomers to functional tetramers (Beernink & Tolan, 1994; Rellos *et al.*, 2000). A space-filling representation of the aldolase monomer with superimposed $C_R(x)$ scores (Figure 6(c)) clearly outlines two surface patches. Figure 6(d) shows the aldolase tetramer (Blom & Sygusch, 1997), a dimer of dimers with two types of dimer interfaces, and highlights the peripheral residues identified by the $C_R(x)$ score. The first interface (subunits A:D and B:C) contains the naturally occurring mutation at position 256 (magenta CPK) (Rellos *et al.*, 2000). Peripheral residues with $C_R(x)$ scores above the 10% background threshold are shown as CPK and colored yellow for the A:D interface. The second interface (red CPKs), between subunits A:B and C:D, is associated with the

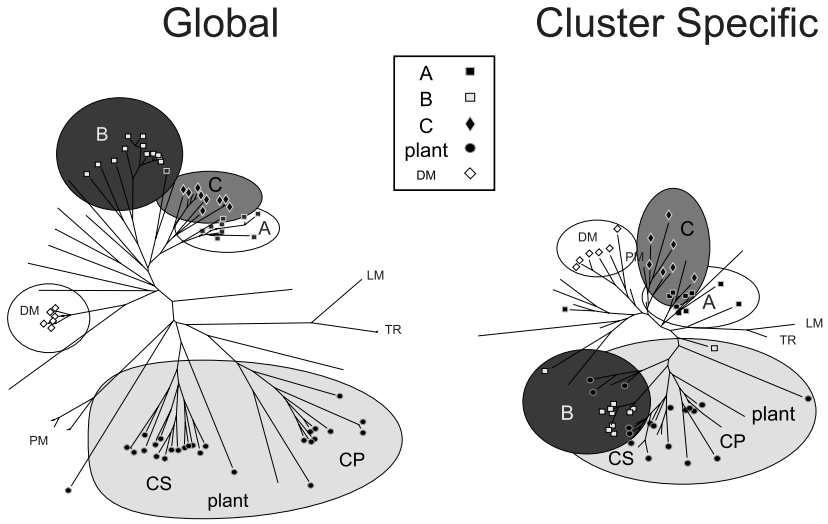

Figure 7. The C-terminal residue cluster of aldolase shows liver isoform-specific shifts in the apparent phylogram. An apparent phylogram was generated with ClustalW and TreeView for the full-length sequences (Global) and residues that make up the C-terminal cluster (cluster specific) based on their similarity deviation scores. Mammalian isoforms of aldolase are marked as A (muscle isoform), B (liver isoform) and C (brain isoform), DM, *Drosophila*; LM, *Leishmania*; TR, *Trypanosomes*; PM, *Plasmodium*. Plant sequences are subdivided into cytosolic (CS) and chloroplast (CP) forms.

mutation in position 128 (cyan CPK) (Beernink & Tolan, 1994).

A moderately conserved non-catalytic residue cluster is located on the backside of the aldolase monomer, relative to the orientation shown in Figure 6(c). The CPK representation (Figure 6(e)) identifies the entrance to the highly conserved active site at the center and the location of residue 128 to the left marks the outer border of the A:B (C:D) interface. To the right of the active site entrance, the white outline (marked AS) identifies a residue cluster with moderate conservation. Peptide mapping studies suggest that this region is involved in the transient, regulatory interaction of aldolase with actin. The corresponding region could not be identified through the use of positional conservation scores (data not shown).

In contrast to ERK2, the similarity deviation scores for aldolase emphasize primarily a residue cluster that differs from those identified by their conservation score (Figure 6(f)). While some overlap exists in sections of the catalytic core, additional high scores can be found near the C-terminal section of aldolase (residues 26 to 32, 72 to 75, 268 to 272, 279 to 289, 297 to 307, 331 to 338), centered on the helix located underneath the flexible C terminus of aldolase. This residue cluster is associated with a naturally occurring mutation in position 337 (shown in CPK). This mutation has been identified in patients with hereditary fructose intolerance and abolishes the equal utilization of F-1-P and FBP (Rellos *et al.*, 1999), unique to the liver isoform of aldolase. Figure 7 shows a comparison of the global phylogram and the cluster-specific phylogram based on the highest-scoring residues in this C-terminal residue cluster (Figure 7). The global phylogram separates the three mammalian aldolases (split further into muscle (A), liver (B) and brain (C) isoforms), plant aldolases (separated into cytosolic and chloroplast forms) as well as *Drosophila* and several parasite aldolase sequences. The cluster-specific phylogram reproduces the same groups of sequences. However, in contrast to the global phylogram, the mammalian liver (B) isoforms are now grouped with cytosolic plant aldolases.

Discussion

In order to evaluate the predictive power of 3D cluster analysis we used two different tests. First, we measured the ability of 3D cluster analysis to identify functionally relevant interfaces, as defined by cocrystal structures, in 35 protein families. We also used this first level of analysis to establish absolute cut-off values (background thresholds) for the *Z*-scores, describing regional conservation ($C_R(x)$) and regional changes in sequence similarity relationships ($S(x)$). Second, we used two proteins to analyze in detail the biochemical relevance of all predicted functional residue clusters. In both tests we evaluated the ability of 3D cluster analysis to achieve our two main goals. First, to improve the accuracy of functional residue cluster prediction, especially for moderately conserved protein-protein interfaces; second, to evaluate whether residue clusters with divergent sequence similarity relationships exist and if they can be correlated with biological functions.

Identification of functional residue clusters based on regional conservation scores

With respect to the detection of known interface residues by 3D cluster analysis, an average of 67 % of residues (63 % for protein-protein interfaces) can be identified by their regional conservation scores ($C_R(x)$) at a background threshold that produces an average of 8 % false positives (37 % detected with a false positive rate of <1 %) (Table 1). Therefore, the majority of interface residues can be predicted with relatively great accuracy. Whilst higher success rates could be achieved for active-site residues, which generally show high sequence conservation,

protein-protein interfaces are more difficult to detect based on sequence conservation patterns. A recent analysis of the level of sequence conservation in stable oligomer interfaces showed that conservation at stable interfaces is higher than the conservation observed for randomly selected surface residue patches, but the conservation signal is generally weak (Jones *et al.*, 2000). One would expect an even lower level of sequence conservation for transient protein-protein interfaces, where the interactions are weaker, and residues have to fulfil the requirement for two types of environments. Residues in transient interfaces have indeed the lowest detection rate at an *E*-score threshold of 10^{-50}. However, a striking feature of 3D cluster analysis is its ability to harvest the information inherent in a large number of more distant homologs, especially for transient interfaces. At an *E*-score threshold of 10^{-20}, 72% of transient interface residues could be detected based on their $C_R(x)$ score, with only 6% anticipated false positives. The same trend, although less pronounced and calculated for a smaller number of examples, is evident for protein-nucleic acid interfaces. Hence, the additional information provided by the up-front inclusion of 3D information allows 3D cluster analysis to make effective use of information inherent in a diverse set of homologous sequences. A comparison of the anticipated rate of false positives for the positional and regional conservation score indicates that $C_R(x)$ is less vulnerable than $C_P(x)$ to errors resulting from the use of ''less than optimal'' alignments. This effect is noteworthy, since multiple sequence alignments are a key element of most evolution-based methods and it explains the beneficial contribution of diverse sequences to 3D cluster analysis.

In the case of ERK2, the improved detection of transient interfaces is evident for the backside residue cluster surrounding the CD domain (Figure 3(b)). The addition of divergent sequences did improve the extent to which this residue cluster is set apart from its surroundings (Figure 3(f)). In contrast, this addition resulted in a reduction of the signal from the positional conservation score (Figure 3(e)). The residue cluster on the backside of ERK2 is centered around and underneath the L16 loop. This loop is of variable length between different MAPKs and shows little conservation by itself, except for a pattern of conserved acidic residues (Asp316 and Asp319 in ERK2) identifying the CD domain (Tanoue *et al.*, 2000). This CD domain serves as a docking site for multiple proteins, including MEK1, MAPK phosphatases (MKPs) and MAPK substrates, and enhances the efficiency of the respective interaction. The specificity with which different MAPKs interact with their signaling partners is still an area of intense investigation. The outline of the conserved residue cluster on the backside of ERK2 is consistent with biochemical data. These data suggest that portions of both the N and C-terminal lobe of ERK2 are involved in the interactions between ERK and MEK (Wilsbacher *et al.*, 1999) or the regulatory phosphatase MKP-3 (Nichols *et al.*, 2000).

The example of the oligomer interfaces in aldolase emphasizes another important feature of the $C_R(x)$ score: its ability to give information on residues under the surface. In contrast to the $C_P(x)$ score, residues above the $C_R(x)$ background threshold that are not part of the direct interface are distributed primarily in the immediate surrounding of the identified interface (Figure 6(b)). Since the calculation of $C_R(x)$ scores includes non-surface residues, the underlying layer of residues is also evaluated. This broader definition of subunit interfaces is supported by the location of naturally occurring mutations, known to interfere with tetramer formation. In addition to residue 128, which takes part in direct interface contacts, two mutations, in position, 256 and 142 (not labeled), are known to affect oligomer formation directly (as opposed to destabilization of the monomer) (Rellos *et al.*, 2000). Both residues are buried and do not make direct interface contacts but are part of the residue cluster identified by 3D cluster analysis.

Sequence clusters with deviating sequence similarity relationships. The *S*(*x*) score

At the outset of this analysis, we asked whether proteins could possess residue clusters for which the global sequence similarity relationships might not adequately reflect evolutionary and functional relationships. This question is of particular importance when global phylogenetic trees are used for the prediction of functional or specificity-conferring regions in a protein. Our analysis of similarity deviation scores ($S(x)$) indicates that such residue clusters do exist and, furthermore, suggests that elevated $S(x)$ scores may indicate regions controlling the specificity of protein functions.

The evaluation of predefined interfaces for their $S(x)$ score shows the highest occurrence of high $S(x)$ scores in active sites, closely followed by transient protein-protein interfaces. A case by case evaluation of residue clusters with high $S(x)$ scores provides more insight into the nature of these clusters. In the case of ERK2, the $S(x)$ score emphasized a portion of the active site (the P1 region) known to confer substrate specificity (Figure 4(a)). A narrow stretch of residues on the backside outlines the CD domain in the L16 loop. The transient interactions of MAPKs with their upstream and downstream partners are known to be highly specific, although the nature of this specificity is still poorly understood. The prediction based on the $S(x)$ score may aid in reconciling the requirement for specificity with the universal nature of the CD motif. A top view in ribbon presentation (Figure 4(d)) illustrates the presence of a four-helix bundle with elevated $S(x)$ scores that makes up the back and core of the C-terminal domain of ERK2. It is tempting to speculate that these residues form a cluster that may be involved in transmitting signal from the binding events at the CD site to the catalytic region

on the front of the protein in a manner that determines the specificity of the outcome. However, no mutational data are available at this point to evaluate this hypothesis.

The only region in ERK2 that stands out primarily on the basis of its $S(x)$ score is the ATP-binding pocket (Figure 4(a) and (c)). As illustrated in Figure 5, the high $S(x)$ score is a reflection of a change in sequence similarities between ERK2 and P38 homologs and plant MAPKs, respectively. The change in the grouping pattern would in fact indicate that this residue cluster represents a strong signature element of mammalian MAPKs. Hence, the identification of key residue clusters for the purpose of classification of proteins with unknown function may be another application of similarity deviation scores.

In the case of aldolase, the $S(x)$ score identifies a residue cluster that (except for a small active-site component) does not coincide with any of the predefined interface categories we evaluated on the above set of 35 protein families. The comparison of the cluster-specific and global phylogram implicates this residue cluster in liver isoform specificity. The regrouping of the liver isoform aldolases with plant cytosolic aldolases in the cluster-specific phylogram is consistent with this assumption. Like the mammalian liver isoform, the cytosolic aldolase from maize preferably catalyzes the cleavage of F-1-P over FBP. Point mutations in the C-terminal tail of maize aldolase have been shown to contribute to this substrate specificity (Berthiaume *et al.*, 1991). Portions of the extreme C terminus of mammalian aldolases have likewise been implicated in the modulation of enzyme specificity (Sygusch *et al.*, 1987). However, the strongest support for the hypothesis that the identified residue cluster is involved in B-isoform specificity, comes from a naturally occurring mutation in human liver aldolase (A337V, CPK in Figure 6(f)), located in the center of the cluster. This mutation, found in patients with hereditary fructose intolerance, abolishes equal utilization of F-1-P and FBP by the liver enzymes (Rellos *et al.*, 1999) and causes a deficiency in the utilization of exogenous fructose. This residue cluster does not stand out based on domain organization or conservation patterns (in contrast to the P1 site or backside interface of ERK2). Aldolase therefore provides an example of how the $S(x)$ score extends the prediction of functionally relevant residue clusters.

Although we do not include mechanistic assumptions in the calculation of $S(x)$ scores, we ask what is the origin of clusters with high similarity deviation scores. Clusters with high $S(x)$ scores are, by definition, characterized by deviations from the averaged sequence similarities of the full-length sequence. We speculate that recombination events that exchange sections of homologous but functionally divergent proteins result in deviations from average sequence similarities. Alternatively, differences in the rates of evolution within the protein sequence and between branches of the phylogenetic tree may account for the observed deviations. A recent study, applying a covarion model to elongation factors, correlates regions in the protein that exhibit divergent rates of evolution among branches of the phylogenetic tree with differences in specificity (Gaucher *et al.*, 2001). Here, we evaluate the $S(x)$ score as a new predictive parameter, and the observed link between functional residue clusters and high $S(x)$ scores suggests that the underlying mechanism contributes to the acquisition of specificity in catalytic and binding events.

Limitations of 3D cluster analysis

Several limitations have to be kept in mind when 3D cluster analysis is applied to homologous sequences. First, 3D cluster analysis identifies functionally relevant residue clusters. Thus, the scores assigned to each residue do not necessarily reflect the relative importance of that residue compared to its immediate neighbors. Examples of this limitation are provided by the conserved positions in the CD domain of ERK2 or the interface point mutants in aldolase. In both cases, the regional conservation score is required to outline the residue cluster of interest. In a next step, the positional conservation scores provide a more accurate measure of the relative importance of individual residues.

A second limitation arises from the assumption of structural similarity within the set of homologous proteins. This assumption is based on the high level of sequence similarity at the *E*-score threshold chosen for our analysis and affects the extent to which more diverse sequences can be included or structures of distant homologs can be used as a reference. To evaluate the effect of variations in the reference structure, we carried out two tests. Using the same set of aligned sequences, we evaluated the ERK2 sequence alignment based on the structure of the closely related P38 MAPK (Wang *et al.*, 1998) (rmsd 1.94 Å). In a second test, we analyzed the aldolase sequence alignment using the monomer structure of the human muscle aldolase as a reference structure (Gamblin *et al.*, 1991) instead of one subunit of the rabbit muscle tetramer (rmsd 0.22 Å). In both cases we obtained the same results with respect to the outlines of the predicted residue clusters, with minor variations for the contribution of individual residues (data not shown). This finding reflects the fact that the probe radius of 10 Å, defining the neighborhood of each residue, acts as a buffer towards minor structural deviations. The probe radius is important with respect to the inclusion of more divergent sequences, determined by the *E*-score threshold. We chose two *E*-scores (10^{-50} and 10^{-20}) at which structural similarity is a relatively safe assumption. However, data sets with a small number of highly homologous sequences may contain insufficient sequence diversity to carry out a successful analysis. Insufficient sequence diversity precludes the

calculation of *Z*-scores (see Materials and Methods) or results in a high percentage of anticipated false positives. A further lowering of the *E*-score or a widening of the probe radius can improve detection in some cases. However, if these parameters are changed, it is crucial that the validity of a predicted residue cluster is evaluated on the basis of a reshuffled alignment model.

In summary, we present a method for the prediction of functionally significant residue clusters in proteins. This method relies on a representative structure and a multiple sequence alignment as input data. Three-dimensional cluster analysis emphasizes the importance of spatially contiguous residue clusters as the sites of functionality within proteins. The inclusion of structural information at the start of the analysis significantly enhances the detection of interfaces that are marked by moderate conservation. This enhancement is particularly pronounced for transient protein-protein interfaces. Functional residue clusters include all residues that contribute to the maintenance of a functional interface. This broader interface definition includes residues that participate directly in protein-protein contacts and others beneath the site that are crucial for the maintenance of the interface.

Our evaluation of changes in regional sequence similarity relationships indicates that a protein with multiple conserved functions may contain residue clusters in which the sequence grouping derived from the full-length sequence does not adequately reflect all aspects of the functional relatedness of the different sequences. We introduce a similarity deviation score to define residue clusters where such deviations exist. This score highlights residue clusters important in conferring specificity within a set of homologous but functionally divergent proteins.

Materials and Methods

Three-dimensional cluster analysis

For each protein family, we select a representative structure, evaluate surface exposure with CCP4-AREAIMOL (version 2.15) (CCP4, 1994) (optional), and identify homologous sequences by a FASTA search (Pearson & Lipman, 1988). Sequences below a set expectation value are aligned using ClustalW (version 1.8, Thompson *et al.*, 1994) to generate a preliminary multiple sequence alignment, which may contain gaps in the sequence representing the reference structure. Next, we remove all positions in the preliminary alignment for which the reference sequence contains a gap. This creates the global, structure-matched alignment denoted by A. A is an $N \times P$ matrix (equation (1)) representing the alignment of all sequences N in all positions P, where P denotes both the length of the alignment and the number of residues in the reference structure:

$$A = [A_{n,p}] \quad n \in [1 \ldots N], p \in [1 \ldots P] \tag{1}$$

The structural environment of a residue is represented by its neighbors in 3D space and is evaluated as outlined in Figure 1. For each residue x in the reference structure, all residues with C^{α} atoms within a set radius (default 10 Å) are considered neighbors ($\eta(x)$) (Step I in Figure 1). The similarity relationships of the structural environment of residue x is represented by the alignment of all positions that are considered structural neighbors of x (Step II in Figure 1). We refer to this alignment as the regional alignment at residue x, denoted $A(x)$. $A(x)$ is a subset of the global alignment, A (equation (2)). All subsequent calculations are based on the comparison of the level of sequence similarity between the regional and global alignment and the results are assigned to residue x:

$$A(x) = [A_{n,p}] \quad n \in [1 \ldots N], p \in [\eta(x)] \tag{2}$$

Next, we construct a similarity matrix for each alignment (Step III in Figure 1). The global similarity matrix (M) is an $N \times N$ matrix containing N^2 sequence similarity terms, denoted m (equation (3)), where each similarity term ($m_{n,n'}$) is a measure of sequence similarity between sequence n and n' in the global alignment (equation (5)). A total of P regional similarity matrices ($M(x)$) are constructed on the basis of the regional alignments, one for each residue in the reference structure:

$$\begin{aligned} M &= [m_{n,n'}] \qquad n \in [1 \ldots N], n' \in [1 \ldots N] \\ M(x) &= [m_{n,n'}(x)] \quad n \in [1 \ldots N], n' \in [1 \ldots N] \end{aligned} \tag{3}$$

As described by Landgraf *et al.* (1999), the global similarity terms ($m_{n,n'}$) are calculated on the basis of the similarity of the full-length sequences in the global alignment A (equation (4)), where $S(A_{n,p}, A_{n'p})$ denotes the substitution score for the replacement of the residue in position p of sequence n with the residue in position p and sequence n'. The substitution score is taken from a positive BLOSUM 62 matrix (Henikoff & Henikoff, 1992). The positive matrix was obtained by subtraction of the lowest (negative) score in the standard BLOSUM 62 matrix from all matrix entries. Values for $m_{n,n'}$ can range from 0 to 1. The alignment of a sequence to itself would produce a score of zero. The regional similarity terms $m_{n,n'}(x)$ are calculated accordingly, using the sequences from the regional alignment $A(x)$:

$$m_{n,n'} = \frac{1}{P} \sum_{p=1}^{P} \frac{s(A_{n,p}, A_{n,p}) - s(A_{n,p}, A_{n',p})}{s(A_{n,p}, A_{n,p})} \tag{4}$$

For each residue in the reference structure, two scores are calculated. The raw regional conservation score $C'_R(x)$ is a measure for the conservation of the structural neighborhood of residue x, compared to the protein as a whole. $C'_R(x)$ is a measure for differences in magnitude between the M and $M(x)$ matrices. For ease of representation the $C'_R(x)$ score was converted to a score from 0 to 1 (initially -1 to 1). Low $C'_R(x)$ scores indicate that the structural neighborhood of residue x shows a higher conservation than the protein as a whole:

$$C'_R(x) = \frac{\left(1 + \sum_{n,n'} \frac{M_{n,n'} - M(x)_{n,n'}}{N^2}\right)}{2} \tag{5}$$

Whilst the $C'_R(x)$ score captures differences in the magnitude of the two matrices, the second score, termed the similarity deviation score, captures differences that are the result of rearrangements of high and low scores within a matrix. The raw similarity deviation score $S'(x)$ (equation (9)) captures this difference by evaluating the correlation between the global and regional similarity

matrix. Scores for $S'(x)$ can range from 0 to 1 (equation (9)) where ρ denotes the correlation coefficient of the two matrices (equation (8)), ranging from −1 to 1, and $\overline{M}$ denotes the mean value for each matrix (equation (7)). The scaling of the $S'(x)$ score from 0 to 1 was done to facilitate the analysis of raw similarity deviation scores for selected data sets:

$$\bar{M} = \frac{\sum_{n,n'} M_{n,n'}}{N^2} \tag{7}$$

$$\rho = \frac{\sum_{n,n'} (M_{n,n'} - \bar{M})(M_{n,n'}(x) - \overline{M(x)})}{\sqrt{\sum_{n,n'} (M_{n,n'} - \bar{M})^2 \sum_{n,n'} (M_{n,n'}(x) - \overline{M(x)})^2}} \tag{8}$$

$$S'(x) = \frac{1 - \rho}{2} \tag{9}$$

To envisage the meaning of the $S'(x)$ score, consider the following example. Given a set of three sequences n_1, n_2 and n_3 that are related in such a way that sequence n_1 is closer to n_2 than n_3, it follows that $m_{n1,n2} < m_{n1,n3}$. However, the local similarity terms might indicate that for the structural neighborhood of residue x, n_1 shows higher similarity to n_3 than n_2, i.e. $m(x)_{n1,n2} > m(x)_{n1,n3}$. This shift in sequence similarity relationships is reflected in a rearrangement of similarity terms in the regional compared to the global similarity matrix. Depending on the average conservation at this residue cluster, there may not be any net change in the value of $C'_R(x)$ but the lack of correlation is reflected in the $S'(x)$ score. For actual data sets, the differences between the regional and global similarity matrices are often the result of both types of changes.

Finally, the raw similarity deviation scores ($S'(x)$) were converted to Z-scores ($S(x)$). This was done by comparison with scores (S) obtained from a regional alignment of randomly picked positions, equal in number to the number of positions in $A(x)$. $\bar{S}$ and $\sigma(S)$ denote the sample mean and standard deviation obtained for scores obtained from 50 independently generated random regional alignments (equation (10)). Random neighbors were chosen equally from the global alignment without regard to their location in the structure but had to adhere to the same optional surface area requirement used for the selection of neighbors in $A(x)$. The $C'_R(x)$ score was converted to a Z-score ($C_R(x)$) similarly. The conversion of both scores to Z-scores standardizes the data sets, which have marked differences in the intrinsic levels of sequence conservation and permits the establishment of absolute Z-score cut-off values (background threshold) for all proteins. Unless otherwise stated, all evaluations in this analysis are based on Z-scores ($C_R(x)$ and $S(x)$):

$$S(x) = \frac{S'(x) - \bar{S}}{\sigma(S)} \tag{10}$$

The two regional scores, describing the properties of the structural neighborhood of residue x, are complemented by a positional conservation score $C_P(x)$ (equation (11)). The positional conservation score does not take the structural environment of residue x into account and merely describes the degree of conservation at position x within the global alignment. With a decreasing probe radius, the $C'_R(x)$ score will converge towards $C_P(x)$:

$$C_P(x) = \frac{1}{N^2 - N} \sum_{\substack{n,n' \\ n \neq n}} \times \frac{s(A_{n,x}, A_{n,x}) - s(A_{n,x}, A_{n',x})}{s(A_{n,x}, A_{n,x})} \quad x \in [1 \ldots P] \tag{11}$$

The resulting scores for all residues were visualized using the reference structure. Clusters of residues with high $S(x)$ scores were further analyzed to evaluate the nature of the shifts in sequence similarity relationships. To this end, high-scoring residues with a C^α distance of less than 6 Å to each other were represented as an ungapped alignment. The value of 6 Å was determined empirically as optimal for most proteins to enforce a clustering of spatially adjacent residues whilst preventing complete fusion of all residues into one uninformative cluster. The sequence similarity relationships of the residues within the extracted residue clusters were evaluated with ClustalW. For clusters of sufficient size, phylograms were generated with TreeView for visualization purposes.

Evaluation of 35 protein families by 3D cluster analysis

The criteria for the selection of proteins for this analysis were the availability of a crystal structure of a complex, identifying the interface between two proteins or the protein and one of its ligands, and the availability of a sufficient number of homologous sequences. Antibody-antigen complexes were not used for the analysis. Homologous sequence for each protein family were compiled at E-score thresholds of 10^{-50} and 10^{-20}, resulting in data sets ranging from 31 to 630 sequences. Data sets that did not have sufficient sequence diversity to calculate Z-scores at the default probe radius of 10 Å were eliminated. Highly redundant data sets result in the selection of randomly assembled neighborhoods consisting only of fully conserved neighbors with a standard deviation of zero, thus preventing the calculation of Z-scores (equation (10)). The following structures were used for the analysis: 1ad0, 1bfn, 1mem, 1eai, 1d7r, 1e61, 1ogs, 1am4, 1blx, 1buh, 1fin, 1ak4, 2pcc, 1efp, 1itb, 1c1j, 1wq1, 1dfj, 1 vol, 1a7k, 1avx, 2uug, 1gdt, 1uaa, 1ser, 1axr and 1ncf.

For the evaluation of interface residues, the exposed surface area for each residue was calculated separately for the protein chain under analysis and the appropriate complex structure using AREAIMOL. In cases where a protein has more than one type of interface, e.g. a small ligand-binding site and an oligomer interface, the interfaces were considered independently. Residues were considered interface residues if the exposed surface area was reduced by more than 30 % upon complex formation. For protein-protein complexes, the proteins were evaluated independently if they met the above criterion. The scores obtained for each protein family were compared with those obtained from a reshuffled alignment. For the reshuffling, each sequence in the alignment was reshuffled independently, maintaining the relative amino acid composition of each sequence.

Acknowledgments

We thank Matteo Pellegrini, Edward Marcotte, Rob Grothe and Parag Mallick for helpful discussions during

the development of this method. We thank Gary Kleiger for discussions and for providing a list of non-redundant protein complex structures. We thank Lukasz Salwinski for his assistance with computational questions. This work was supported by the DOE (DE-FC03-87ER-60615) and NIH (GM 31299), an NIH-NRSA fellowship to R.L. and Swiss national fellowship to I.X.

References

Anderson, N. G., Maller, J. L., Tonks, N. K. & Sturgill, T. W. (1990). Requirement for integration of signals from two distinct phosphorylation pathways for activation of MAP kinase. *Nature,* **343**, 651-653.

Beernink, P. T. & Tolan, D. R. (1994). Subunit interface mutants of rabbit muscle aldolase form active dimers. *Protein Sci.* **3**, 1383-1391.

Berthiaume, L., Loisel, T. P. & Sygusch, J. (1991). Carboxyl terminus region modulates catalytic activity of recombinant maize aldolase. *J. Biol. Chem.* **266**, 17099-17105.

Blenis, J. (1993). Signal transduction via the MAP kinases: proceed at your own RSK. *Proc. Natl Acad. Sci. USA,* **90**, 5889-5892.

Blom, N. & Sygusch, J. (1997). Product binding and role of the C-terminal region in class I D-fructose 1,6-bisphosphate aldolase [letter]. *Nature Struct. Biol.* **4**, 36-39.

Blumer, K. J. & Johnson, G. L. (1994). Diversity in function and regulation of MAP kinase pathways. *Trends Biochem. Sci.* **19**, 236-240.

Boulton, T. G. & Cobb, M. H. (1991). Identification of multiple extracellular signal-regulated kinases (ERKs) with antipeptide antibodies. *Cell Regul.* **2**, 357-371.

Bucher, P. & Bairoch, A. (1994). A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Cell Regul.* **2**, 53-61.

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171-178.

Collaborative Computational Project Number 4 (1994). The CCP4 program suite: programs for protein crystallography. *Acta Crystallog. sect. D,* **50**, 760-763.

Clarke, F. M. & Masters, C. J. (1975). On the association of glycolytic enzymes with structural proteins of skeletal muscle. *Biochim. Biophys. Acta,* **381**, 37-46.

Clarke, F. M. & Morton, D. J. (1976). Aldolase binding to actin-containing filaments. Formation of paracrystals. *Biochem. J.* **159**, 797-798.

Clarke, F. M. & Morton, D. J. (1982). Glycolytic enzyme binding in fetal brain-the role of actin. *Biochem. Biophys. Res. Commun.* **109**, 388-393.

Clarke, F. M., Stephan, P., Huxham, G., Hamilton, D. & Morton, D. J. (1984). Metabolic dependence of glycolytic enzyme binding in rat and sheep heart. *Eur. J. Biochem.* **138**, 643-649.

Cooper, S. J., Leonard, G. A., McSweeney, S. M., Thompson, A. W., Naismith, J. H., Qamar, S., Plater, A., Berry, A. & Hunter, W. N. (1996). The crystal structure of a class II fructose-1,6-bisphosphate aldolase shows a novel binuclear metal-binding active site embedded in a familiar fold. *Structure,* **4**, 1303-1315.

Crews, C. M., Alessandrini, A. & Erikson, R. L. (1992). The primary structure of MEK, a protein kinase that phosphorylates the ERK gene product. *Science,* **258**, 478-480.

Dalby, A., Dauter, Z. & Littlechild, J. A. (1999). Crystal structure of human muscle aldolase complexed with fructose 1,6-bisphosphate: mechanistic implications. *Protein Sci.* **8**, 291-297.

Davis, R. J. (1993). The mitogen-activated protein kinase signal transduction pathway. *J. Biol. Chem.* **268**, 14553-14556.

Gamblin, S. J., Cooper, B., Millar, J. R., Davies, G. J., Littlechild, J. A. & Watson, H. C. (1990). The crystal structure of human muscle aldolase at 3.0 Å resolution. *FEBS Letters,* **262**, 282-286 [published erratum appears in *FEBS Letters*, 1990, **264**, 159].

Gamblin, S. J., Davies, G. J., Grimes, J. M., Jackson, R. M., Littlechild, J. A. & Watson, H. C. (1991). Activity and specificity of human aldolases. *J. Mol. Biol.* **219**, 573-576.

Gaucher, E. A., Miyamoto, M. M. & Benner, S. A. (2001). Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl Acad. Sci. USA,* **98**, 548-552.

Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565-6572.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.

Horecker, B. L., Tsolas, O. & Lai, C. Y. (1972). Aldolases. In *The Enzymes* (Boyer, P. D., ed.), 3rd edit., vol. 7, Academic Press, New York.

Jones, S. & Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121-132.

Jones, S. & Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133-143.

Jones, S., Marin, A. & Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77-82.

Landgraf, R., Fischer, D. & Eisenberg, D. S. (1999). Analysis of heregulin symmetry by evolutionary tracing. *Protein Eng.* **12**, 943-951.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358.

Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325-337.

Mansour, S. J., Matten, W. T., Hermann, A. S., Candia, J. M., Rong, S., Fukasawa, K., Vande Woude, G. F. & Ahn, N. G. (1994). Transformation of mammalian cells by constitutively active MAP kinase kinase. *Science,* **265**, 966-970.

Nichols, A., Camps, M., Gillieron, C., Chabert, C., Brunet, A., Wilsbacher, J., Cobb, M., Pouyssegur, J., Shaw, J. P. & Arkinstall, S. (2000). Substrate recognition domains within extracellular-signal regulated kinase mediate binding and catalytic activation of MAP kinase phosphatase-3. *J. Biol. Chem.* **275**, 24613-24621.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444-2448.

Penhoet, E., Rajkumar, T. & Rutter, W. J. (1966). Multiple forms of fructose diphosphate aldolase in

mammalian tissues. *Proc. Natl Acad. Sci. USA,* **56**, 1275-1282.

Penhoet, E. E., Kochman, M. & Rutter, W. J. (1969). Ioslation of fructose diphosphate aldolases A, B, and C. *Biochemistry,* **8**, 4391-4395.

Rellos, P., Ali, M., Vidailhet, M., Sygusch, J. & Cox, T. M. (1999). Alteration of substrate specificity by a naturally-occurring aldolase B mutation (Ala337 → Val) in fructose intolerance. *Biochem. J.* **340**, 321-327.

Rellos, P., Sygusch, J. & Cox, T. M. (2000). Expression, purification, and characterization of natural mutants of human aldolase B. Role of quaternary structure in catalysis. *J. Biol. Chem.* **275**, 1145-1151.

Robinson, M. J., Harkins, P. C., Zhang, J., Baer, R., Haycock, J. W., Cobb, M. H. & Goldsmith, E. J. (1996). Mutation of position 52 in ERK2 creates a nonproductive binding mode for adenosine 5′-triphosphate. *Biochemistry,* **35**, 5641-5646.

Schlessinger, J. (1994). SH2/SH3 signaling proteins. *Curr. Opin. Genet. Dev.* **4**, 25-30.

Sygusch, J., Beaudry, D. & Allaire, M. (1987). Molecular architecture of rabbit skeletal muscle aldolase at 2.7-Å resolution. *Proc. Natl Acad. Sci. USA,* **84**, 7846-7850.

Tanoue, T., Adachi, M., Moriguchi, T. & Nishida, E. (2000). A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nature Cell Biol.* **2**, 110-116.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53-64.

Walsh, T. P., Masters, C. J., Morton, D. J. & Clarke, F. M. (1981). The reversible binding of glycolytic enzymes in ovine skeletal muscle in response to tetanic stimulation. *Biochim. Biophys. Acta,* **675**, 29-39.

Wang, Z., Canagarajah, B. J., Boehm, J. C., Kassisa, S., Cobb, M. H., Young, P. R., Abdel-Meguid, S., Adams, J. L. & Goldsmith, E. J. (1998). Structural basis of inhibitor selectivity in MAP kinases. *Structure,* **6**, 1117-1128.

Wilsbacher, J. L., Goldsmith, E. J. & Cobb, M. H. (1999). Phosphorylation of MAP kinases by MAP/ERK involves multiple regions of MAP kinases. *J. Biol. Chem.* **274**, 16988-16994.

Xu, D., Lin, S. L. & Nussinov, R. (1997). Protein binding *versus* protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.* **265**, 68-84.

Zheng, C. F. & Guan, K. L. (1993). Dephosphorylation and inactivation of the mitogen-activated protein kinase by a mitogen-induced Thr/Tyr protein phosphatase. *J. Biol. Chem.* **268**, 16116-16119.

Edited by J. Thornton

(Received 25 September 2000; received in revised form 29 January 2001; accepted 30 January 2001)