© 1995 Nature Publishing Group http://www.nature.com/nsmb

# A method to predict functional residues in proteins

Georg Casari, Chris Sander and Alfonso Valencia

EMBL-Heidelberg D-69012 Heidelberg, Germany

**The biological activity of a protein typically depends on the presence of a small number of functional residues. Identifying these residues from the amino acid sequences alone would be useful. Classically, strictly conserved residues are predicted to be functional but often conservation patterns are more complicated. Here, we present a novel method that exploits such patterns for the prediction of functional residues. The method uses a simple but powerful representation of entire proteins, as well as sequence residues as vectors in a generalised 'sequence space'. Projection of these vectors onto a lower-dimensional space reveals groups of residues specific for particular subfamilies that are predicted to be directly involved in protein function. Based on the method we present testable predictions for sets of functional residues in SH2 domains and in the conserved box of cyclins.**

Biological sequence data are accumulating rapidly as a result of advanced sequencing technology and concerted genome projects. The probability that a new protein can be classified as a member of a sequence family is already near 50%[1]. The more members of a family are known, the more we begin to learn about the evolutionary constraints that conserve residues or their properties at particular sequence positions. Evolutionary constraints are imposed by requirements of three-dimensional structure and of biological function. In general, functional requirements are known to be more pronounced in terms of residue identities than structural constraints: completely conserved residues in a dispersed protein family usually have a direct role in function. For example, the conserved Ser-His-Asp triad of serine proteinases performs the key steps in catalysis; similarly, the conserved Asn of the Asn-Lys-X-Asp motif of G-domains makes a specific pair of hydrogen bonds to the guanine base of bound GTP or GDP. Given a multiple sequence alignment, it is generally straightforward to spot the most conserved residues and predict their involvement in function. Mutation of such residues typically causes loss of protein function.

Sequence conservation is less obvious for residues that modulate the specificity of biological function. Such residues change as a protein evolves to satisfy modified functional constraints, while the basic biochemical mechanism and the overall three-dimensional fold remain unaltered. For example, the difference in peptide cleavage specificity between trypsin and chymotrypsin is achieved by residues of the required chemical type in the specificity pocket near the active site. Evolutionary changes in specificity tend to occur in jumps, that is residues that determine specificity are conserved within a subfamily of proteins, but differ between subfamilies. This stepwise behaviour is consistent with an evolutionary scenario in which functional requirements change rather sharply with a change in specificity and remain constant thereafter. In general recognition of functional subfamilies by their characteristic residues is not easily done by mere inspection of a multiple sequence alignment.

Here we describe a straightforward and powerful new method to identify residues that are likely to be responsible for functional differences between protein subfamilies. The approach requires only a multiple sequence alignment as input and provides an experimentally testable prediction in the form of likely functional residues. The analysis is illustrated with examples in which there is a proven correlation between prediction and experiment. We show testable predictions derived from this type of sequence analysis for two protein families of biological interest where experimental evidence is incomplete or not yet available.

The method is based on analysis of protein multiple sequence alignments which can be generated using well-tested algorithms[2]. Typically, the sequences are retrieved by scanning the protein sequence databases for homologues of a search sequence. Ideally, the family of homologues has several divergent members rather than almost identical sequences. As a rule of thumb, one includes pairs with fewer than 50% identical residues. The quality of the multiple alignment is crucial, as all further results depend on it.

© 1995 Nature Publishing Group http://www.nature.com/nsmb

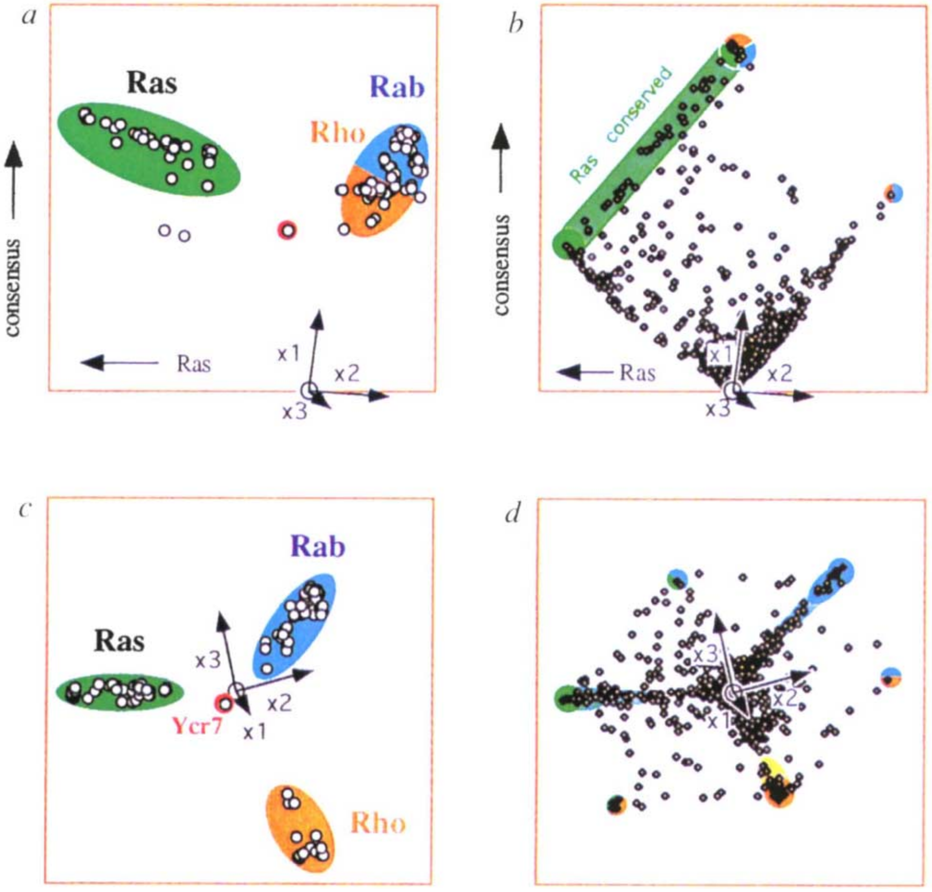


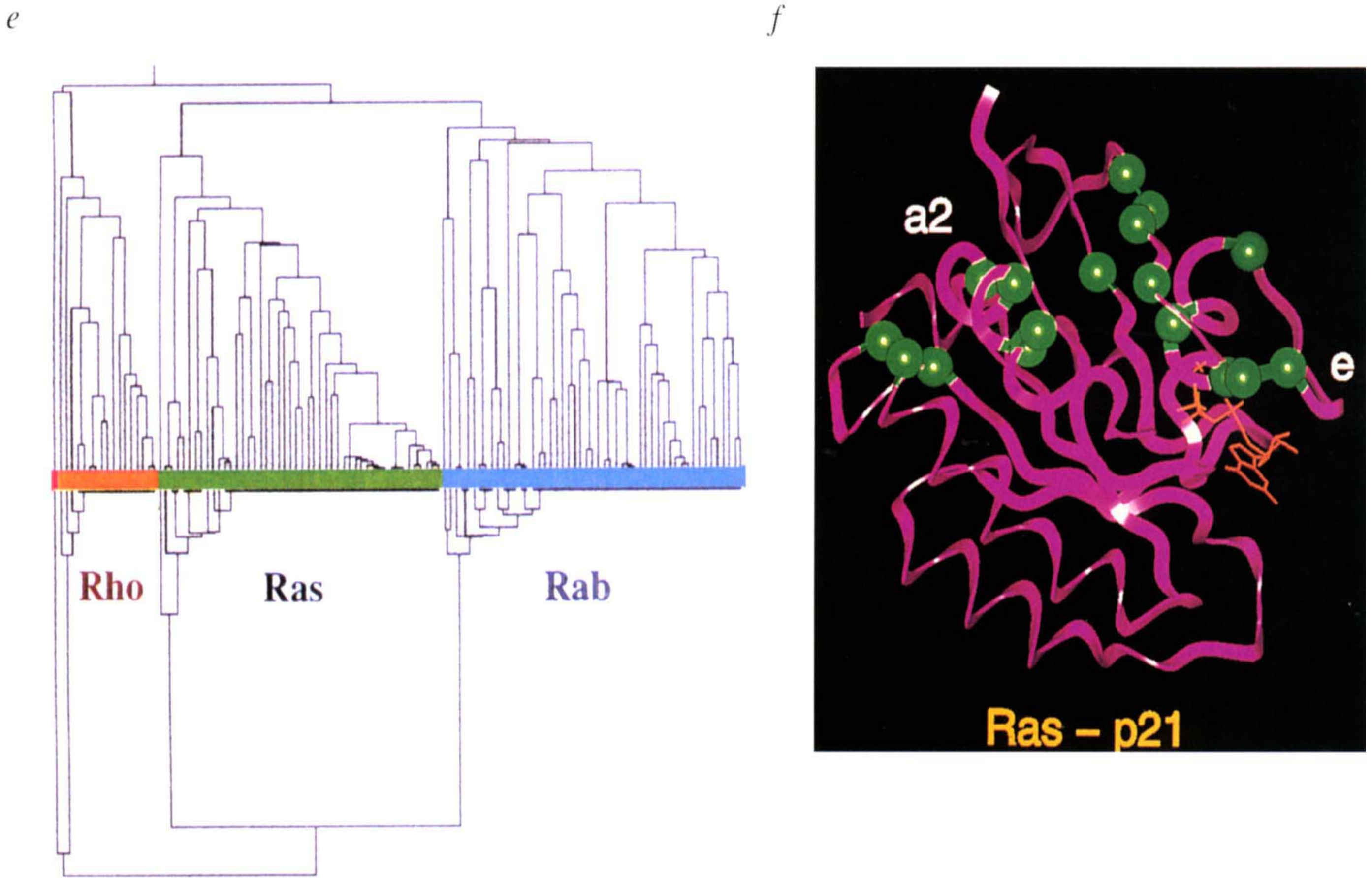
**Fig. 1** Sequence space analysis of the Ras-Rab-Rho superfamily. Two projections of the Ras-Rab-Rho superfamily defined by the three principal axes with largest eigenvalues, *x*1, *x*2 and *x*3. Proteins (open circles) are shown in the left hand plots (*a*,*c*), positions of single sequence residues (diamonds) projected onto the same planes are shown in corresponding right hand plots (*b*,*d*). Projection of *a*, proteins and *b*, residues, onto a plane containing the direction of the consensus sequence pattern for the entire Ras-Rab-Rho superfamily (vertical) and a Ras specific direction (horizontal). In (*a*), proteins more representative of the Ras subfamily are farther to the left. In (*b*), single residues completely conserved in all three families occupy the extreme top corner (tricoloured), all of which are involved in GTP-binding or hydrolysis. Residues specific for the Ras subfamily form a corner in the Ras-specific direction (left, green) and all residues conserved in the Ras family (specific or not) occupy the upper left edge (light green band) in this representation. *c*,*d*, Projection onto a plane containing the same Ras specific direction (horizontal) and a discriminating direction for Rho (vertical). In this representation all three protein families form separate clusters. Ycr7, a yeast protein (red), is not a member of any of these clusters and most likely the only representative of a new functional family. Single residues occupying the remote corners in direction of Ras, Rab and Rho are specific for the corresponding families and predicted to participate in their specific function. Specific residues shared between two families form corners in directions between those specific for the single families (highlighted in two colours). *e*, *next page*, Family tree of Ras-like proteins. The upper tree has been obtained from comparisons of complete sequences. The lower tree is obtained using only the subset of specific residues picked as corners from *d*; (black diamonds). Classification into subfamilies becomes clearer and unambiguous. *f*, *next page*, Ras-specific residues predicted to be crucial for function (highlighted green) in the known structure of human Ras[22]. They occupy the confined surface region of switch II (around α-helix 2; marked α2) and switch I (effector loop e) known to interact with GAP and the nucleotide exchange factor. *g*, *next page*, The conservation pattern in the switch II region evident in a selected subset of sequences illustrates the essence of the method in a simple case. Some conserved residues involved in GTP binding are indicated by a box in line 'conserved'. Specific residues for the subgroups of Ras, Rab and Rho as obtained from this analysis are marked with asterisks in the corresponding lines.

## A vectorial representation

The novelty of the approach comes from a mathematically convenient representation that allows grouping of protein families and identification of characteristic sequence patterns in a unified fashion. We represent each sequence as a vector point in a multi-dimensional space, (sequence space), with residue positions and residue types as the basic dimensions. The formalism of principal component analysis[3] can then be applied to determine the directions in sequence space most strongly populated by the proteins in the family. Although the visual representation of proteins in this subspace appears similar to previous methods using multivariate statistics for low-dimensional representation of protein families[4], the underlying mathematical concepts are very different. These fundamental differences enable us not only to define the protein subfamilies, as do other methods, but also at the same time, to trace the principal components back to the individual residues and positions that are characteristic of the different subfamilies.

The geometric origin of sequence space as defined here is the central point of reference. Relative to the origin, both direction and length of the vectors have a biological meaning. Directions in sequence space represent specific sequence patterns (profiles), combinations of specific residue types at specific sequence positions. The directions of the principal axes can be interpreted as the sequence patterns that best discriminate between members of the protein family. Typically, the direction of the first principal axis (largest eigenvalue) corresponds to the consensus pattern of the entire family. Preferred directions in this space reveal which residue types at which sequence positions best distinguish a subfamily. The resulting concept is simple: proteins of a particular subfamily as well as residues characteristic for the function of this subfamily point to the same direction in sequence space.

npg © 1995 Nature Publishing Group http://www.nature.com/nsmb



*g*

```
                 --beta 2--  --beta3---       -alpha 2-  -beta 4   ---- alpha 3 ----
rash_human       EDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRV
rapa_human       EDSYRKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFALVYSITAQSTFNDLQDLREQILRV
rsr1_yeast       EDSYRKTIEIDNKVFDLEILDTAGIAQFTAMRELYIKSGMGFLLVYSVTDRQSLEELMELREQVLRI
Ras specific     * ***              *          **  *   *                           *
sec4_yeast       IDFKIKTVDINGKKVKLQLWDTAGQERFRTITTAYYRGAMGIILVYDVTDERTFTNIKQWFKTVNEH
rab5_canfa       AAFLTQTVCLDDTTVKFEIWDTAGQERYHSLAPMYYRGAQAAIVVYDITNEESFARAKNWVKELQRQ
rab7_canfa       ADFLTKEVMVDDRLVTMQIWDTAGQERFQSLGVAFYRGADCCVLVYDVNSVKSFDNLNNWREEFLIQ
Rab specific       *              *        *       *  *        *            *
rhoa_human       FENYVADIEVDGKQVELALWDTAGQEDYDRLRPLSYPDTDVILMCFSIDSPDSLENIPEKWTPEVKH
rac1_human       FDNYSANVMVDSKPVNLGLWDTAGQEDYDRLRPLSYPQTDVFLICFSLVSPASYENVRAKWFPEVRH
```

The lengths of vectors represent the degree of conservation. A protein more distant from the origin (Fig. 1*a*) is more representative in that it contains a larger fraction of residues characteristic for this direction (pattern) in the subspace. Similarly, the most strongly conserved residues take the most distant positions (Fig. 1*b*) and form edges of the region occupied by all residues. Residues conserved in only one subfamily form the distant edge in the direction of this subfamily. Residues conserved in two subfamilies occupy the corner where the two corresponding edges meet (Fig. 1*d*). Clear clusters of residues on these corners and edges can reflect strong evolutionary selection and their member residues are predicted to be directly involved in function. In short, by representing both single residues and proteins in the principal dimensions of sequence space, protein subfamilies are evident as clusters and characteristic residues as corners and edges.

Biologically interesting directions in sequence space can also be defined by exploiting *a priori* experimental knowledge, rather than by principal component analysis of sequence data alone. In this case functional axes that separate proteins according to their known function are defined manually assigning +1 for proteins of

© 1995 Nature Publishing Group http://www.nature.com/nsmb

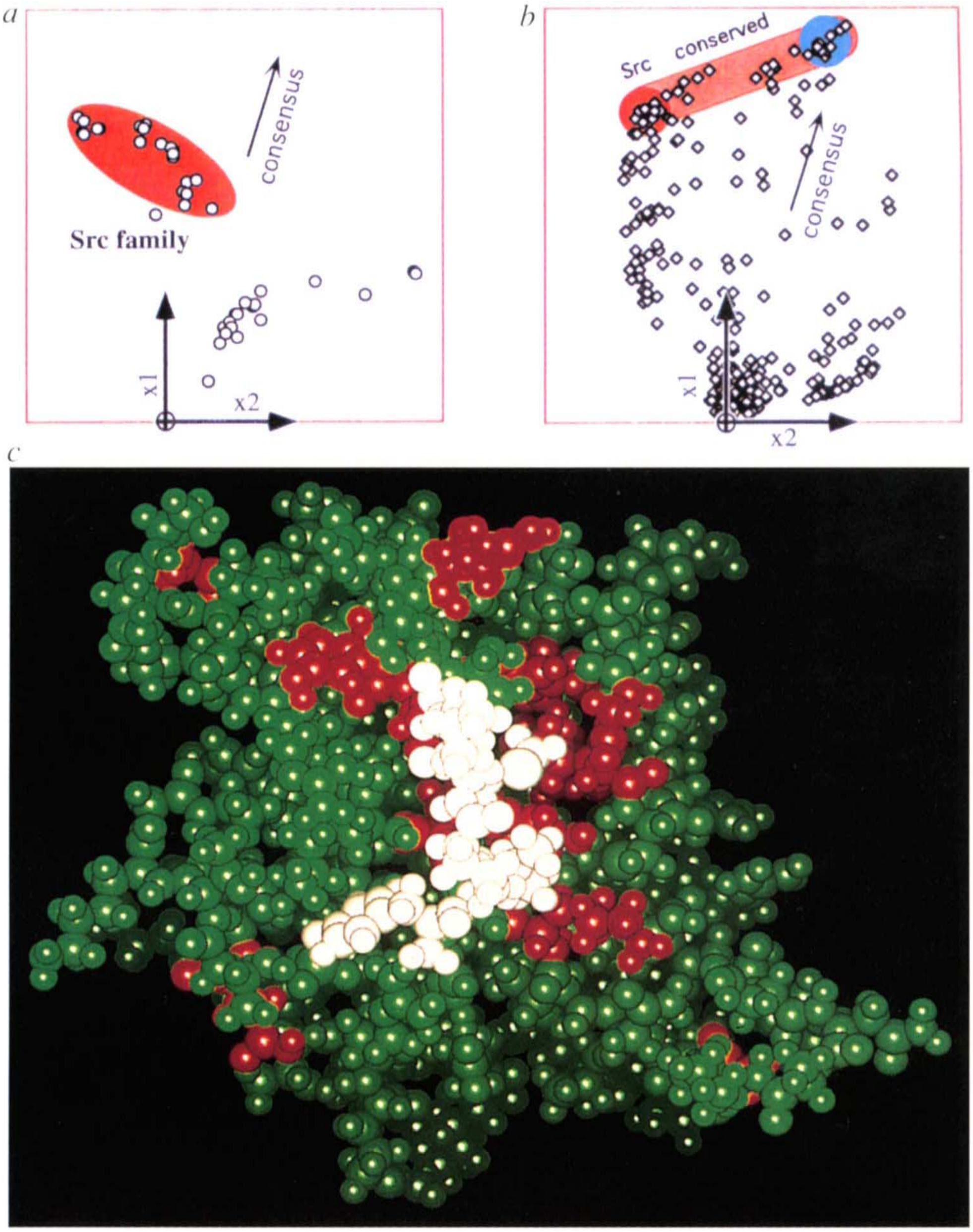


**Fig. 2** Analysis of SH2 domain sequences for Src-like specificity. *a*, SH2 domains projected onto the plane defined by the first two principle axes (each circle represents an entire domain). The plane in sequence space defined by the first two principle axes contains a direction specific for the common consensus of all SH2s (arrow) and a direction specific for Src homologous sequences (Src, Fyn, Lyn, Yes; ellipsoid). *b*, Projection of residues onto the same plane, used as the basis for prediction of Src-specific residues (each diamond represents a residue type at a specific position). Single residues conserved in all SH2 domains are identified at the extreme point of the consensus direction (blue circle: W7, R14, L21, G29, F31, L32, R34, S36, S46, H61, I64, F80 and L86 in 1SHA), while residues unique to the cluster of Src homologous kinases are identified at the extreme point of the Src specific direction (red circle: L22, N26, K40, D49, L67, G71, C98 and C106). Residues predicted to be functional occupy the edge of Src conservation between these extreme points (light red band: E37, L45, G56, K60, Y62, G70,Y73, I74, R77, Y90, G96, 97L, L10). The underlined residues are known to make specific contacts in crystal structures of SH2 domains complexed with peptide substrates (only one known contact residue is missed here)[18]. The other seven residues constitute a genuine prediction of functional residues. *c*, Mapping of predicted functional residues (red) onto the known 3D structure of the v-Src SH2 domain, used as basis for independent verification of the prediction. Evidence in support of the prediction comes from the fact that most of the residues not yet known to be functional extend the surface patch of those residues already known to be in contact with the bound phosphotyrosyl peptide (white).

the functional class and -1 for nonfunctional proteins. Projection of residues onto these external axes suggests the most likely residues involved in the specific function that was externally defined (not shown).

## Functional residues in Ras-like proteins

We first illustrate the method by applying it to the Ras-Rab-Rho family of small GTPases, for which much experimental information about functional residues is available permitting direct evaluation of the accuracy of prediction. This family has a simple phylogenetic tree with three clear subfamilies of distinct biological function (Fig. 1*e*, top): Ras proteins involved in signal transduction, including the human p21 Ras protooncogene protein; Rab proteins involved in specific targeting of vesicles within the cell; and Rho proteins involved in organisation of the cytoskeleton[5,6]. All of these proteins bind and hydrolyse GTP. On hydrolysis, a major conformational change takes place in the region of α-helix 2 (switch II) that acts as a signal for other cellular proteins[7,8].

Applying sequence space analysis to a multiple sequence alignment of 116 member proteins (entry 5p21.hssp in database HSSP[9]), the first three principal axes (*x*1, *x*2, and *x*3 in Fig. 1) define the most revealing directions. Let us concentrate on the directions defined by lines from the origin to a protein or a residue. The first axis ($x_1$) points in the direction of a sequence pattern common to all Ras-Rab-Rho proteins. Residues located in this direction are common to all Ras-Rab-Rho

© 1995 Nature Publishing Group http://www.nature.com/nsmb

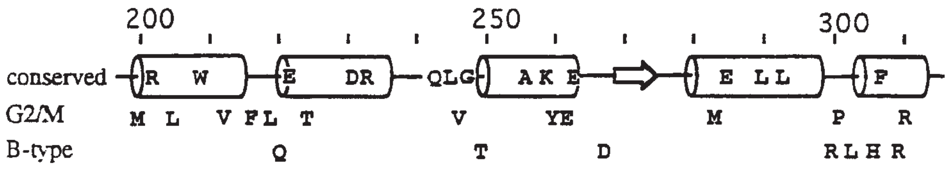


**Fig. 3** Genuine predictions of functional residues for cyclins. Schematic representation of the cyclin box with conserved, G2/M specific and B-type specific residues. Predicted α-helices[23] are depicted as rods, β-strands as arrows. Sequence positions are according to the human sequence of B1 cyclin (cgb1_human). Conserved residues are inserted into the schema (R202, W208, E221, D231, R232, Q245, L246, G248, A265, K267, E270, E286, L290, L293 and F305). The second row marks positions of G2/M specific residues ( M201, L205, V212, F216, L218, T222, V247, Y258, E259, Y277, M285, P300 and R308) and in the third row B-cyclin specific residues are indicated (Q220, T250, D268, L310, H304 and R307).

proteins (tricoloured corner at the top of Fig. 1*b*). These completely conserved residues map to the GTP/GDP binding site[6] in the three-dimensional structure of Ras p21 and almost all of them are known to be involved either in nucleotide binding or in catalysis.

The second axis (*x*2) distinguishes the Ras subfamily from other members of the superfamily and defines a Ras-specific direction in sequence space (Fig. 1*a,c*). Residues conserved in the Ras family form an edge (green band in Fig. 1*b*) that runs from the cluster of Ras-Rab-Rho conserved residues (top in Fig. 1*b*) in the direction of Ras (left in Fig. 1*b*). The more closely the vector of a residue points in the direction of the Ras cluster, the more specific it is for Ras proteins. The high conservation of all residues along this edge is most likely the result of strong evolutionary selection and these residues are predicted to be involved in function.

The direction of the third principal axis (*x*3) primarily separates Rho proteins from Ras and Rab. Rab, Ras and Rho subfamilies are best distinguished in this plane of axes *x*2 and *x*3 (Fig. 1*c*) and form clear clusters. Single residues represented in this plane (Fig. 1*d*) are spread out according to their specificity for either Ras, Rab or Rho. Residues at the end of bisectors, where two edges meet, have dual specificity. For example, those at the lower left in Fig. 2*d* (green/orange Tyr 40, Arg 68, Pro 34) are conserved in Ras and Rho, but not in Rab proteins. They are likely to be responsible for a functional aspect, defined at the single residue level, common to both Rho and Ras. An equivalent example is position 71. A Ser at this position is specific for Rho proteins. Rab and Ras share a Tyr, a case of dual specificity (Fig 1*g*).

Supporting evidence for our interpretation of sequence space comes from an alternative analytical method and from experiment. When an evolutionary tree is constructed using only the subset of residues at the three specificity corners (Ras, Rab, Rho corners of Fig. 1*d*), the structure of the tree simplifies greatly (Fig. 1*e*). The principal evolutionary event associated with the simplified tree is the functional differentiation of a primordial protein into the Ras, Rab and Rho functionality. This leads to an alternative definition of specificity residues as the subset resulting in the sharpest evolutionary tree. The sequence space analysis method efficiently determines such a residue subset in a simple one-step procedure.

Experimental confirmation of functional predictions can best be illustrated by mapping the Ras-specific residues onto the three-dimensional structure of Ras p21. Remarkably, these residues map to the switch I and switch II regions (Fig. 1*f*) known to be important for Ras function[7]. This regions make interactions with GAP, the GTPase activating protein, and the nucleotide exchange factor, that induces reloading of GTP. Switch I covers the end of helix α1 and the effector loop where we identify Ala 18, Thr 20, Gln 25, Asp 33 and Pro 34. Switch II participates in signal readout: in elongation factor Tu, a remote homologue of Ras p21[10], a strong conformational change in helix α2 (switch II) as a result of GTP hydrolysis alters an exposed surface patch that strongly affects the interaction with other domains of EF-Tu[11–13]. By close analogy, the Ras GTPase switch is read out by its molecular partners interacting with a similar surface patch. The patch is predicted here to include residues Met 67, Arg 68, Tyr71 and Gly 75, on helix α2; Val 103, Lys 104, Pro 110 on helix α3; and Glu 37, Ser 39, Tyr 40, Arg 41, Leu 56 on strands β2 and β3.

## Functional residues in SH2 domains

A case illustrating the difficulties of visual analysis of multiple alignments is the group of Src homology 2 domains (SH2). This family has many subgroups and few sequences in each subgroup. SH2 domains bind phosphotyrosine containing proteins and peptides with high specificity and many play a key role in regulatory processes (for example, those in protein kinases). In spite of several solved 3D structures, the full extent of the specificity pocket is not yet known[14].

Analysing SH2 domains from kinases, phospholipases and spectrins using the sequence space approach we find that the first two principal axes are the most informative. (The alignment of 52 SH2 sequences, created with the multiple alignment program Maxhom[15], can be obtained from the authors on request.) As before, the first principal axis describes the consensus of all sequences, while the second axis distinguishes Src-type receptor kinase sequences (Fig. 2*a*), a group known to have the same sequence specificity for phosphotyrosine peptides[16]. In addition to the completely conserved residues, those on an edge extending from the completely conserved cluster (top in Fig. 2*b*) to the cluster of Src-specific residues (left in Fig. 2*b*) are predicted to be functionally important. To prevent overprediction, we exclude from consideration the cluster of Src-specific residues, as all sequences in this cluster have high pairwise sequence similarity.

Mapping the predicted residues onto the known crystal structure of the *Src*-SH2 low affinity phosphotyrosyl peptide complex[17] confirms the accuracy of the prediction for some of the residues and leaves others as a genuine prediction. Completely conserved residues either participate in a small hydrophobic core that stabilises

© 1995 Nature Publishing Group http://www.nature.com/nsmb

the SH2 fold or map to the binding pocket for phosphotyrosine. The predicted thirteen specific residues mostly cover a nearby cleft. Remarkably, six of these overlap with residues known to make specific contacts with the bound peptide[18], only one of which is not identified. Seven more are predicted to be involved in contacts providing extra specificity on larger peptides. Residues identified in the complex and those predicted here (Fig. 2*c*) define the specific binding site. This hypothesis can be tested by mutagenesis experiments of binding-site residues predicted to alter peptide binding or change specificity.

## Functional residues in cyclins

The predictive power of the method can be tested most stringently in a case where no 3D structure is known. Cyclins are involved in control of cell cycle progression. Different cyclin types specifically control entry to different cell cycle states[19]. Sequence space analysis of the cyclin box, a homologous region in all cyclins (110 residues, 49 sequences), identifies, as before, a set of completely conserved residues, a set of residues functionally specific for B-type cyclins, and a set of residues on a connecting edge. The latter set is predicted to be G2/M specific, as they are exclusively found in cyclins involved in

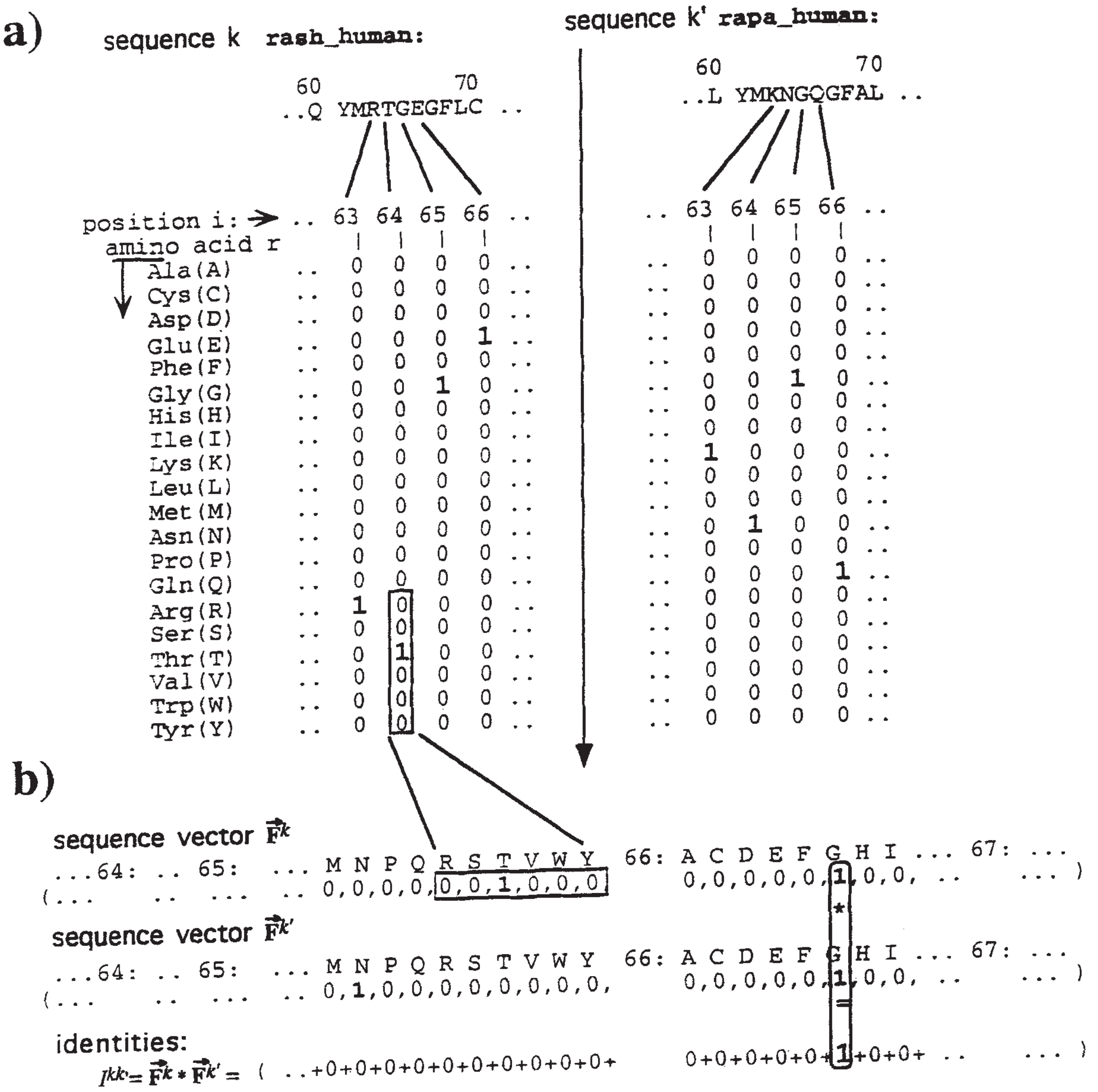


**Fig. 4** Representation of two protein sequences as vectors in sequence space: *a*, Illustration of the translation of two sequences *k* and *k'* (top) into tables, with entries for each residue type at each position (profiles). The number 1 is entered for the residue type that actually occurs at a particular position in the sequence, the number 0 for all other residue types. Rearranging the entries of the table results in sequence vectors as shown in (*b*). These vectors define the location of the corresponding sequences in sequence space. *b*, The number of identities $I^{kk'}$ between sequence *k* and *k'* can be calculated as the vector product of the sequence vectors $\vec{F}^k$ and $\vec{F}^{k'}$. A box highlights an identity between sequences *k* and *k'* corresponding to their common G residue at position 66.

© 1995 Nature Publishing Group http://www.nature.com/nsmb

control of transition from G2 phase to mitosis (B-type and B-like cyclins from plants and fungi). We predict that introducing these residues into a different cyclin box (for example, that of cyclin A) should be sufficient to switch this domain to G2/M specificity.

The prediction can be refined somewhat by mapping residues onto the predicted secondary structure (Fig. 3). Completely conserved residues are likely to be involved in forming the structural core (hydrophobic residues near the centres of the predicted helices) or involved in crucial functional interactions (hydrophilic residues in loops or helix ends), possibly with cdc2. G2/M specific residues in the predicted loops and helix ends are predicted to form a separate surface patch that forms interactions to specifically control the G2/M transition.

## Scope and utility

The sequence space method developed here predicts functional residues by exploiting evolutionary information in a set of related sequences; the 'fossil record' of evolution under selective pressure. As shown in the control examples the new method picks up subtle patterns of conservation and is capable of accurate predictions. The predictions have two levels. On the first level, residues identified as completely conserved and residues identified as subfamily specific are predicted to be involved in some (initially unspecified) aspect of function. The implication for experiment is that point mutations in these residues are predicted to lead to a strong phenotpe. On the second level, in favourable cases, particular functions can be assigned (in the predictive sense) to particular residues. Such more detailed predictions require some detailed knowledge of the biological functions common to all proteins in the family (these are assigned to the completely conserved residues) and/or of the functions specific to proteins in particular subfamilies (these are assigned to the residues specific for each subfamily). In the control example, residues common to the entire Ras Rab Rho family would be predicted to be involve in nucleotide binding, while Ras specific residues would be predicted to be involved in interactions with specific GTPase activating proteins, nucleotide exchange factors, or downstream effectors. For the SH2 domains of protein kinases, for which the most conserved residues are known to be involved in phosphotyrosine binding, the subfamily-specific residues are predicted to be involved in sequence-specific binding of the flanking peptide.

Like any predictive method, there is a margin of error in the predictions. The precise level of error will become apparent as the method is used in practice to plan specific experiments. The likelihood of error is higher when sequence information is sparse. The method works best when ample sequences well dispersed in the subfamilies are available to triangulate sequence space. In simple cases, the results of the method agree with intuitive analysis of conserved residues by inspection of multiple alignments ('sequence gazing'), in the identification of completely conserved residues. However, the method may be superior to sequence gazing in the identification of more subtle sequence patterns that are difficult to pick out by eye and require labour intensive inspection of both multiple sequence alignments and family trees.

The analytical and predictive power of the method stems from the introduction of a conceptually novel view of sequence diversity, with complementary representation of both proteins and residues in the same mathematical space. The practical advantage of the method becomes apparent when a large sequence family is analysed and/or when conservation patterns are subtle. We anticipate that the sequence space method will be useful to molecular biologists as a tool to aid prediction and classification of functional residues and for planning targeted residue-specific functional experiments.

## Methods

The sequence vectors $\vec{\mathbf{F}}$ used in this approach are analogous to conventional sequence profiles derived from multiple alignments. Just as profiles give a tabular summary of the amino acid content at each position in an alignment, a sequence vector consists of 1s and 0s, depending only on whether a particular residue type is present at a sequence position or not (Fig. 4). When rearranged as a row vector $\vec{\mathbf{F}}^k$, the kth protein sequence corresponds to a point in 20*l*-dimensional space, where *l* is the length of the sequence alignment and *k* is the index for the protein. The alignment **F** of *n* sequences is a matrix of *n* rows of length 20*l* , each row holding the components of a single sequence vector.

$$\mathbf{F} = \left| \begin{array}{c} \vec{F}^{1} \\ \\ \vec{F}^{k} \\ \\ \vec{F}^{n} \end{array} \right|$$

The number of identities $C^{kk}$ between sequences *k* and *k'* can be expressed as the inner product of the sequence vectors (Fig. 4*b*).

$$C^{kk'} = \vec{F} \bullet \vec{F}^{k} = \sum_{i,r} F_{ir}^{\ k} F_{ir}^{\ k}$$

A comparison matrix **C** with the number of identities for all pairs of sequences can thus be expressed as the matrix product between alignment **F** and its transpose $\mathbf{F}^T$:

$$\mathbf{C} = \mathbf{F}\,\mathbf{F}^{T}$$

All possible sequences of length up to *l* residues can be represented as points in the 20 *l* -dimensional sequence space. Members of a protein family typically populate only a small region of this space, as they are similar in both length and sequence. As a result, the main features of a family can be described in a subspace of a smaller number of dimensions. The subspace most suitable for the description of a particular family can be found by solving an eigenvalue problem: the principle axes defining the subspace are the eigenvectors corresponding to the largest eigenvalues $l_p$ of the comparison matrix **C** (ref. 3). The relations embed $\mathbf{C}\vec{u}_p = l_p\vec{u}_p^{20}$ define the principle axes $\vec{u}_p$, and the coordinate $x_p^k$ of protein *k* in dimension *p* is $x_p^k$ = (sqrt. $\lambda_p u_p^k$). Representation of the family members in a lower-dimensional subspace, for example, as points in a two-dimensional graph, illustrates the main similarity relationships (Fig. 1). Subfamilies are revealed as clusters, analogous to major branches in a tree representation. Higher dimensions, in order of decreasing eigenvalues, describe increasingly finer details in the similarity relationships, analogous to sub-branches in a tree.

The conceptual advantage of the novel vector representation of sequences becomes evident when vectors of individual residues (columns in alignment **F**), rather than those of entire sequences (rows in **F**), are projected along the principle directions. The principle axes defined as sequence patterns $\vec{v}^p$ with components $v^p$ can be obtained as $\mathbf{F}^T\vec{u}_p\lambda_p^{-1/2} = \vec{v}^p$ where $\vec{u}_p$ is an eigenvector and $\lambda_p^{i,r}$ the corresponding eigenvalue of comparison matrix **C**. Coordinates $y_p^{i,r}$ of residue *r* at position *i* in the sequence are $y_{i,r}^p$ = sqrt.( $\lambda_p v_{p,r}^{p)}$ (ref. 3). In this way individual residues can be

© 1995 Nature Publishing Group http://www.nature.com/nsmb

placed in the same space as the full-length protein sequences. This unified representation emphasizes links between sequence subfamilies and their corresponding characteristic residues.

The algorithm has been implemented in a computer program that reads a multiple alignment of sequences (alignment formats MSF[21] or HSSP[9]) and represents proteins as well as individual residues as points in two-dimensional subspaces of sequence space (Fig. 1). The axes *x*1, *x*2, *x*3. of the graphs correspond to the first few principle axes. Different choices of axes, *x*1 and *x*2 in Fig. 1*a, b*; or *x*2 and *x*3 in Fig. 1*c, d* bring out different specificity aspects. When coupled to an interactive graphics program, the user can explore sequence families and their specific residue patterns and can identify likely functional residues.

The executable computer program called Sequence Space and a graphics program to view the results called Scatter will be made available for academic users by Internet from ftp.EMBL-Heidelberg.de and www.EMBL-Heidelberg.de.

**Received 22 August; accepted 24 December 1994.**

**Acknowledgements**
We thank D. Thomas and L. Holm for critical reading of the manuscript and U. Surana for helpful comments on cyclins. This work was supported by an EMBO postdoctoral fellowship (G.C.) and a bioinformatics grant from the German Minstry of Research (BMFT).

**1.** Koonin, E.V., Bork, P. & Sander, C. Yeast Chromosome III: New Gene Functions *EMBO J.* **13**, 493–503 (1994).

**2.** Doolittle, R.F. & Feng, D.-F. Nearest neighbor procedure for relating progressively aligned amino acid sequences *Meths. Enzymol.* **183**, 659–669 (1990).

**3.** Lebart, L., Morineau, A. & Warwick, K.M. *Multivariate Descriptive Statistical Analysis* (John Wiley & Sons, New York; 1984).

**4.** Higgins, D.G. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets *Comput. appl. Biosci.* **8**, 15–22 (1992).

**5.** Bourne, H.R., Sanders, D.A. & McCormick, F. The GTPase superfamily: a conserved switch for diverse cell functions *Nature* **348**, 125-132 (1990).

**6.** Valencia, A., Chardin, P., Wittinghofer, A. & Sander, C. The *ras* protein family: evolutionary tree and role of conserved amino acids *Biochemistry* **30**, 4637–4648 (1991).

**7.** Milburn, M.V., *et al.* Molecular switch for signal transduction: Structural differences between active and inactive forms of protooncogenic ras proteins *Science* **247**, 939–945 (1990).

**8.** Stouten, P.F.W., Sander, C., Wittinghofer, A. & Valencia, A. How does the switch II region of G-domains work? *FEBS Letts* **320**, 1–6 (1993).

**9.** Sander, C. & Schneider, R. The HSSP data base of protein structure-sequence alignments *Nucleic. Acids Res.* **21**, 3105–3109 (1993).

**10.** Valencia, A., Kjeldgaard, M., Pai, E.F. & Sander, C. GTPase domains of *ras* p21 oncogene protein and elongation factor *Tu*: Analysis of three-dimensional structures, sequence families, and functional sites *Proc. natn. Acad. Sci. U.S.A.* **88**, 5443–5447 (1991).

**11.** Kjeldgaard, M., Nissen, P., Thirup, S. & Nyborg, J. The crystal structure of elongation factor EF-Tu from Thermus aquaticus in the GTP conformation *Structure* **1**, 35–50 (1993).

**12.** Kjeldgaard, M. & Nyborg, J. Refined structure of elongation factor Tu from *Escherichia coli J. molec. Biol.* **223**, 721–742 (1992).

**13.** Berchtold, H., *et al.* Crystal structure of active elongation factor Tu reveals major domain rearrangements *Nature* **365**, 126–132 (1993).

**14.** Pawson, T. & Gish, G. SH2 and SH3 domains: From structure to function *Cell* **71**, 359–362 (1992).

**15.** Sander, C. & Schneider, R. Database of homology-derived structures and the structural meaning of sequence alignment *Proteins* **9**, 56–68 (1991).

**16.** Songyang, Z., *et al.* SH2 domains recognize specific phosphopeptide sequences *Cell* **72**, 767–778 (1993).

**17.** Waksman, G., *et al.* Crystal structure of the phosphotyrosine recognition domain SH2 of v-Src complexed with tyrosine phosphorylated peptides *Nature* **385**, 646–653 (1992).

**18.** Birge, R.B. & Hanafusa, H. Closing in on SH2 Specificity *Science* **262**, 1522–1524 (1993).

**19.** Lew, D.J. & Reed, S.I. A proliferation of cyclins *Trends cell Biol.* **2**, *77–81 (1992).*

**20.** Press W.H., Teukolsky S.A., Vetterling WT. & Flannery B.P., in *Numerical Recipes in C* 456–493 (Cambridge University Press, Cambridge, 1992).

**21.** Devereux, J., Haeberli, P. & Smithies, O. A comprehensive set of sequence analysis programs for the VAX *Nucl. Acids Res.* **12**, 387–395 (1984).

**22.** Pai, E.F., *et al.* Refined structure of the triphosphate conformation of h-Ras P21 at 1.35 Angstroms resolution: Implications for the mechanism of GTP hydrolysis *EMBO J.* **9**, 2351–2359 (1990).

**23.** Rost, B., Sander, C. & Schneider, R. PHD - an automatic mail server for protein secondary structure prediction *CABIOS* **10**, 53–60 (1994).