CMLS, Cell. mol. life sci. 53 (1997) 621–645
1420-682X/97/080621-25 $ 1.50 + 0.20/0


CMLS Cellular and Molecular Life Sciences

# Review

# Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives

**I. Callebaut[a], G. Labesse[a], P. Durand[a], A. Poupon[a], L. Canard[a], J. Chomilier[a], B. Henrissat[b] and J. P. Mornon[a,*]**

[a]Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS URA 09, UP6/UP7, Case 115, 4 place Jussieu, F-75252 Paris Cedex 05 (France), Fax +33 1 44 27 37 85, e-mail: Isabelle.Callebaut@lmcp.jussieu.fr
[b]Centre de Recherches sur les Macromolécules Végétales**, CNRS, BP53, F-38041 Grenoble Cedex 9 (France)

**Abstract.** Ten years after the idea of hydrophobic cluster analysis (HCA) was conceived and first published, theoretical and practical experience has shown this unconventional method of protein sequence analysis to be particularly efficient and sensitive, especially with families of sequences sharing low levels of sequence identity. This extreme sensitivity has made it possible to predict the functions of genes whose sequence similarities are hardly if at all detectable by current one-dimensional (1D) methods alone, and offers a new way to explore the enormous amount of data generated by genome sequencing. HCA also provides original tools to understand fundamental features of protein stability and folding. Since the last review of HCA published in 1990 [1], significant improvements have been made and several new facets have been addressed. Here we wish to update and summarize this information.

**Key words.** HCA; sequence analysis; helical net; protein folding.

**Abbreviations.** 1D, 2D, 3D, 4D for one, two, three and four dimensions, respectively.

## Introduction

Genome sequencing provides an enormous amount of protein sequences. However, many remain structurally and functionally uncharacterized although, in principle, they could be related to sequences belonging to a same structural and functional family. Indeed, it is now expected that the entire pool of independent folds does not exceed more than about one thousand [2], i.e. between two and three times the independent folds presently known at the atomic level. Each sequence should code for one of the folds included in this limited set, each fold being compatible with a large number of sequences [3, 4] even if the duration of protein evolution has been far too short for all the possibilities to emerge. Consequently, any given sequence has a considerable probability of sharing similar fold(s) with proteins coding for structures already characterized. The challenge is to recognize these folds and/or functions from sequences only, as the extent of sequence divergence often

* Corresponding author.
** Affiliated with the Joseph Fourier University.

obscures their relationships. This calls for the development of increasingly sensitive methods of sequence analysis.

Current programs of database screening (the most commonly used being BLAST [5] and FASTA [6]) are generally highly efficient but their 'automatic' use is limited to the case of reasonable similarity, i.e. above a 'twilight' zone that contains both relevant and irrelevant hits and where biologically significant similarity is not necessarily statistically significant in itself [7]. Although the limits of this zone are fuzzy, depending on the sequence analysed, its upper limit is estimated at about 25–30% sequence identity over a sufficient length. Below this threshold, and especially at high sequence divergence, 'linear' (1D) methods are generally unable to distinguish sequence similarities reflecting three-dimensional (3D) relationships from background noise. The emergence of efficient motif and profile searches [8] has made it possible partly to overcome these limitations. However, these methods necessitate the prior recognition of a family with members sufficiently divergent to be able to identify striking family features, therefore excluding the case where a query sequence does not show detectable similarity with other proteins. These methods also aim at families that evolved under sufficient pressure (functional or structural) to have maintained recognizable 1D features. In the case of high divergence, seeking authentic 3D relationships therefore involves the use of a combination of methods tested with various parameters as well as the consideration of correlative factors, the success of these investigations often depending on the user's experience and training.

Hydrophobic cluster analysis (HCA) [1, 9], starting from a two-dimensional (2D) helical representation of protein sequences, can help overcome the above limitations of 'linear' analyses. Indeed, it combines the comparison of sequences and that of the protein secondary structures statistically centered on hydrophobic clusters [10]. The 'twilight' zone can therefore be efficiently analysed by putting the observed sequence similarities in the 2D context of the protein. Moreover, the particular interest of the method also lies in the fact that it is not always dependent on the prior detection of 1D similarities by current methods, thereby revealing structural relationships even when no sequence conservation has been highlighted. Numerous applications as well as theoretical studies (see below) have now established the efficiency of this approach.

Furthermore, by focusing on the residues forming the hydrophobic core of proteins, HCA appears to be a powerful tool to investigate the basis of protein stability and folding. Indeed, it now appears that many properties related to protein folding can be highlighted and characterized through HCA, the $\alpha$-helical clustering of hydrophobic amino acids matching these properties very closely and to an unexpected degree. In particular, the distribution of hydrophobic amino acids along the sequences of globular domains is not random and reveals signatures tightly associated with the different folds selected by nature. These data could therefore be used as structural predictive tools. Among others, a universal determinant of globular domain folding detected by HCA is the very limited number of hydrophobic positions in the sequences (about 5 to 10% of the total amino acids) necessary to allow the existence of stable folds (unpublished results).

## Principles of HCA

### HCA plots as support of sequence information

Helical wheels as well as 2D projections of transmembrane segments have long been included in figures illustrating amphipathic and transmembrane $\alpha$-helices. However, apart from this limited and local use of unconventional 1D sequence representation, helical 2D representations of amino acid sequences did not really begin until 1968 with a paper by Dunhill visualizing chemical and structural properties of $\alpha$-helices through vertical diagrams [11]. In the 1970s, the Russian school of Poutschino improved these representations by detailing further the importance of hydrophobic amino acid clusters for secondary and tertiary structure predictions. In particular, V. I. Lim was an inventive pioneer in this field (see e.g. refs 12, 13). More than ten years elapsed before a further advance occurred with the development of the Hydrophobic Cluster Analysis method [9]. Two decisive features were introduced and enhanced the interest and efficiency of the approach. Probably for the first time, the actual goal was the development of a self-consistent method to compare and analyse bidimensional representations of amino acid sequences of any secondary structure. Second, a deliberate choice was made to design a method adapted to the powerful ability of the human eye-brain system to recognize, decipher and associate complex images with disparate and often elusive biological information.

To this end, HCA uses a limited number of four amino acid symbols instead of the current one-letter code, intermediate between the large number of twenty symbols used in the helical representation of Dunhill and the lack of symbols in the Russian school (fig. 1). We have chosen to highlight only four particular amino acids:

– a star (★) for proline which confers the greatest constraint to the polypeptide chain,

– a diamond (◆) for glycine which in contrast confers the largest freedom to the chain,

– squares with a point (◘) or not (□) for serine and threonine, respectively, as these two small polar amino

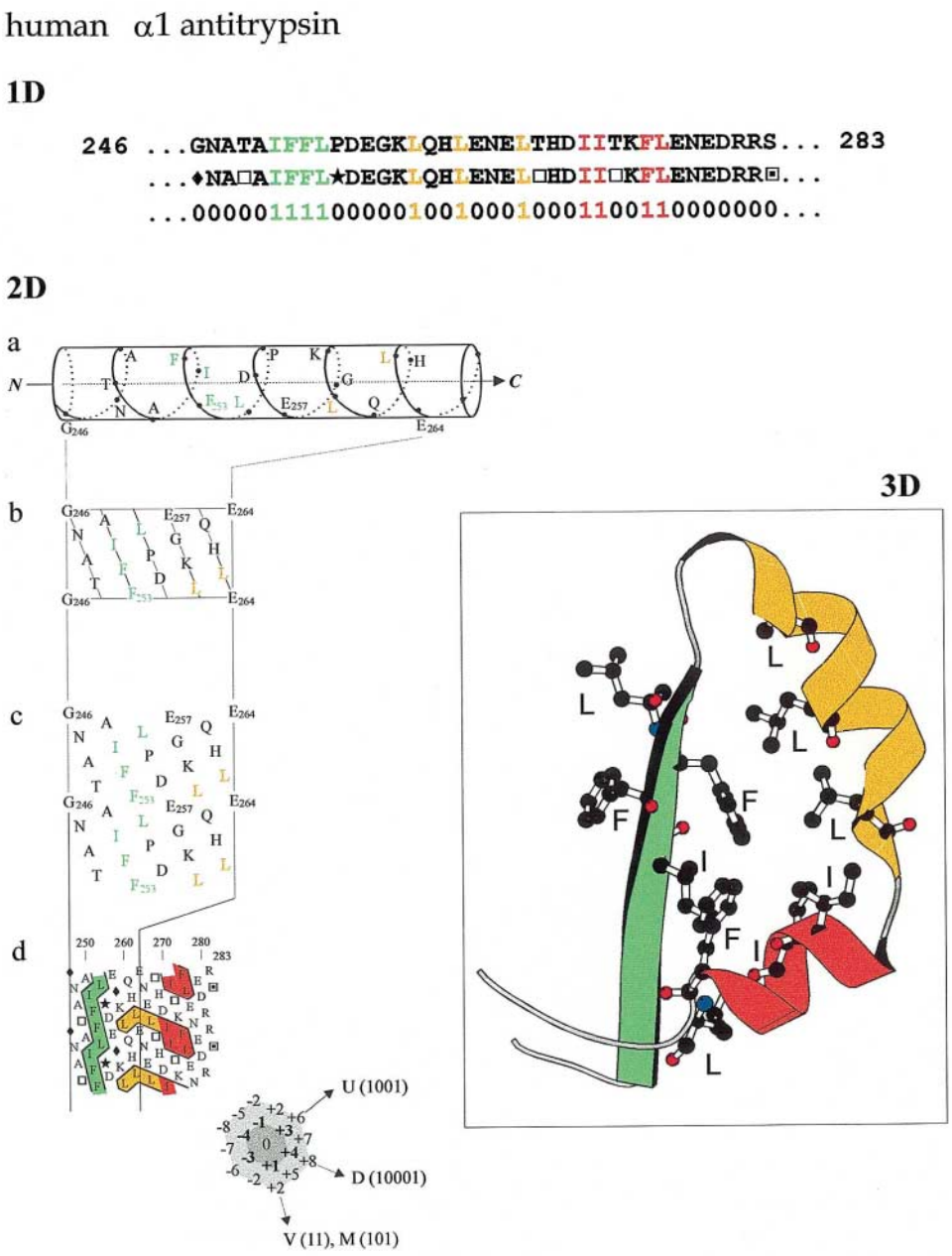


Figure 1. Hydrophobic cluster analysis: between sequence and structure. Principles of the 2D HCA diagram. The linear sequence (1D) of a segment of human $\alpha$1-antitrypsin is shown on top with hydrophobic amino acids coloured (first line). The same sequence is also represented using the HCA code for the amino acids G, P, T and S (second line). The direct translation into a two-state code (1 = hydrophobic and 0 = nonhydrophobic) is represented. The sequence is reported on an $\alpha$-helix displayed on a cylinder (*a*). The cylinder is then cut parallel to its axis and unrolled to a bidimensional diagram (*b*) which is compacted and duplicated in order to restore the full environment that each amino acid has on the $\alpha$-helix representation (*c*). Hydrophobic amino acids are not randomly distributed but tend to form clusters (contoured in *d*). These clusters were statistically shown to correspond to the internal faces of regular secondary structures [10], as strikingly illustrated here by the corresponding experimental three-dimensional structure (3D insert). The vertical cluster, shaded green, is associated with a $\beta$-strand whereas the horizontal one, shaded yellow and orange, corresponds to $\alpha$-helices, illustrating that the general shape of the cluster is indicative of the type of secondary structure. Moreover, the rupture in the orientation of the horizontal cluster (two different colors) can here be related to the split of the $\alpha$-helix into two parts in the 3D structure. Sequence stretches separating clusters correspond to loops (or hinges between domains, especially if they are of considerable length, see fig. 8). The 2D structure of a protein can therefore easily be deciphered through examination of the horizontal texture of the plot. A hydrophobic cluster is defined by contiguous hydrophobic residues on the helical representation. Consequently, for an $\alpha$-helix, two hydrophobic amino acids belong to distinct clusters if they are separated by at least four nonhydrophobic residues (coded 0) or a proline ('connectivity distance' of 4). The three fundamental axes of the 2D $\alpha$-helical plot are associated with four 'basic' clusters: V(11), M(101), U(1001) and D(10001) (see text).

acids can mask their polarity through H-bonding with the carbonyls of the main chain, particularly within helices.

Probably another decisive feature of the HCA bidimensional representation lies in the horizontal orientation of the plots, first suggested by C. Gaboriaud, rather than the vertical one used previously. At least for the western scientist, horizontal reading from left to right is more natural and leads to the best interaction between the eye-brain 'computer' and the bidimensional representation.

At this level, the helical net representation of amino acid sequences can be considered only as a convenient stratagem, like any other arbitrary representation. However, as shown hereafter, it actually fits well with fundamental principles that govern protein folding, reveals important information hidden in the sequences, and offers an efficient way to overcome several of the limitations of classical analysis.

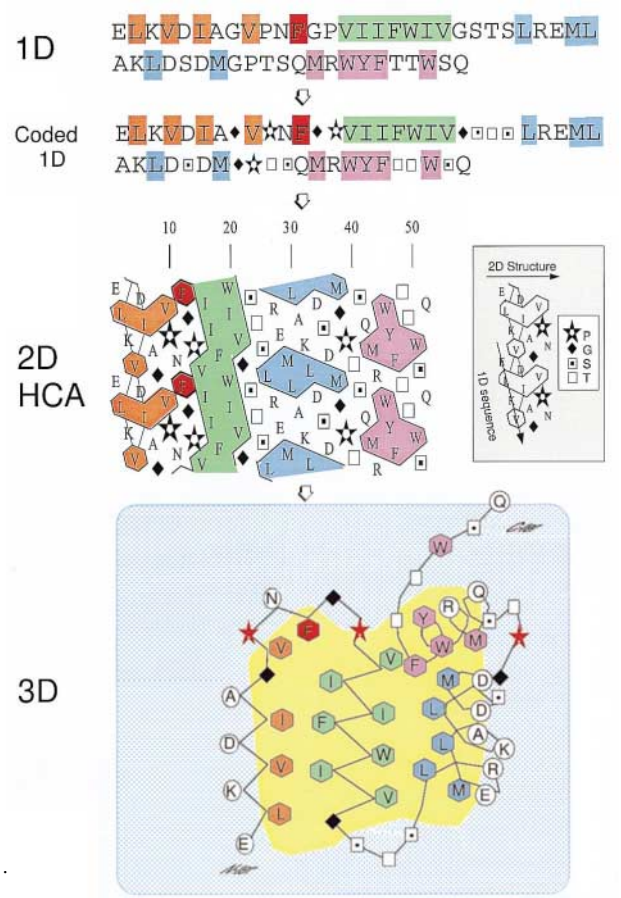


Figure 2. Correspondence between 1D sequence, 2D HCA plot and 3D organization of a schematic globular domain in which hydrophobic amino acids are coloured according to their presence in the buried faces of successive regular secondary structures. From N-ter to C-ter, there is an edge $\beta$-strand, an internal $\beta$-strand, an amphiphilic $\alpha$-helix and a short internal helix, respectively. They summarize the current local structures of globular domains, the least frequent one being the internal helix. The yellow area indicates the limits of the hydrophobic core of the domain immersed in a water surrounding.

## HCA between sequence (1D) and structure (3D)

It is now well established that many biological phenomena are governed by the fundamental dichotomic behaviour of hydrophobic and hydrophilic molecular entities. The general architecture of protein globular domains provides a clear example (fig. 2). Indeed, mostly hydrophobic amino acids dominate the internal core, whereas the mostly hydrophilic amino acids lying on the protein surface protect the core from solvent (water and ions). This partition is mainly due to the fast escape of hydrophobic amino acids from water under the pressure of entropic and enthalpic forces. As a result, the hydrophobic amino acids tend to cluster into a stable and compact structure which is typical of a particular fold. Consequently, the distribution of amino acids along sequences belonging to a single structural family should respect at least several constraints and should give rise to hydrophobic signatures characteristic of the fold. Within globular domains, the mean content in hydrophobic amino acids is close to 1/3 (using VILFMWY alphabet as the list of hydrophobic residues). Hydrophobic amino acids are known to be favoured within the internal faces of regular secondary structures ($\alpha$-helices and $\beta$-strands) and disfavoured within the main irregular secondary structures (loops).

These features explain why, using appropriate representations that take into account local proximities, clusters of hydrophobic amino acids are statistically found in close correspondence with the internal faces of regular secondary structures [10]. Fig. 2 illustrates these correspondences for a schematic globular protein.

To a first approximation, the twenty amino acids are separated into two main classes: the hydrophobic ones and the others which are hydrophilic and/or indifferent to their environment. HCA experience and extensive calculations lead to the following partition:

– strong hydrophobic amino acids VILF which are the driving forces for the constitution of hydrophobic internal faces of secondary structures.

– moderately hydrophobic amino acids MWY, each with its own particularities. M as well as W and Y can accept exposed or semi-exposed situations. These latter two often mediate intermolecular interactions, and Y often favours loops.

These two classes are grouped together to contour hydrophobic clusters in HCA plots and are coded 1 in the HCA nomenclature.

– proline (P), which often breaks secondary structures, is currently considered as a cluster breaker.

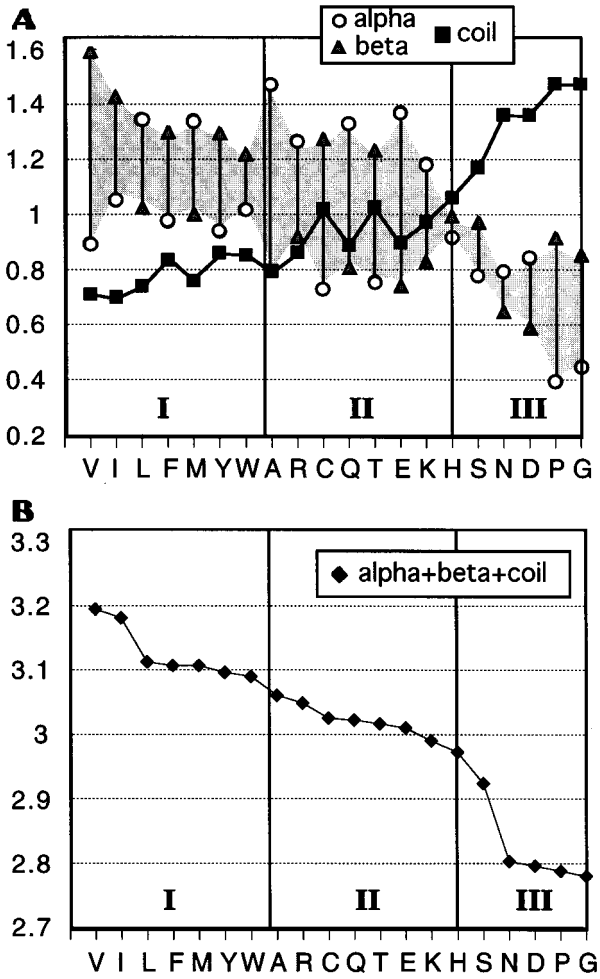


Figure 3. (*A*) Amino acid propensities for the $\alpha$, $\beta$ and coil states, calculated from a nonredundant version of the Protein Data Bank (25% list of the PDB-select version of November, 1996 (553 chains) [14] in which no two proteins have more than 25% sequence identity). The calculated propensities correspond to the percentage of each amino acid observed in each of the three states divided by the percentage of the total amino acids observed in the corresponding states. Secondary structure assignments were performed using the P-SEA algorithm [15]. (*B*) Sums of the three propensities for each amino acid. Three regions can be delineated from these two figures. The first one (I) in which both the $\alpha$ and $\beta$ states are favoured relative to the coil state corresponds to the usual hydrophobic alphabet. The third region (III) typically gathers the loop builders G, P, D, N, S for which the coil state clearly dominates the $\alpha$ and $\beta$ states. The intermediate region (II) contains the other amino acids for which only one of the $\alpha$ or $\beta$ states is preferred to the coil one.

– all the other amino acids are considered in a single class coded 0, although many of them exhibit clear particularities:

– polar amino acids possessing a long aliphatic stem such as arginine (R) or lysine (K) are sometimes observed to subsitute hydrophobic amino acids if their head reaches a polar environment (solvent or ionic pairs).

– asparagine (N) can adopt left helical main chain conformations otherwise principally occupied by glycine (G).

– alanine (A), which clearly favours the formation of $\alpha$-helices, occupies hydrophilic situations as well as hydrophobic ones. Including it in the hydrophobic alphabet leads to a dramatic modification of cluster properties.

– in contrast, cysteine (C), although also adapted to different situations, does not disturb the basis of cluster behaviour. However, its specific role in the formation of disulphide bonds argues against including it in the HCA hydrophobic alphabet.

– threonine (T), a branched amino acid, frequently replaces valine (V) or isoleucine (I) in $\beta$-strands and, like serine (S), can mask its polarity inside $\alpha$-helices; serine can often also mimic the proline ring.

– glutamic acid (E) and especially aspartic acid (D) are the most distant amino acids from VILF in the hydrophobic scale described here, which is similar to many others used elsewhere.

Interestingly, the amino acid classification with strong hydrophobic amino acids (V, I, L, F), the driving forces of regular secondary structures, along with moderately hydrophobic amino acids (M, Y, W) at one extreme, and loop builders (P, G, D, N, S) at the other extreme, is confirmed in a direct experimental observation. Indeed, if the secondary structure propensities of each amino acid to be in the $\alpha$, $\beta$ or coil conformation are added (fig. 3), the resulting index ranges between 3.20 (V) and 2.78 (G) and has the following order: V, I, L, F, M, Y, W, A, R, C, Q, T, E, K, H, S, N, D, P, G. Clearly, the main driving forces of regular secondary structures are the four strong hydrophobic amino acids V, I, L, F and the main driving forces of loops are P, G, D, N, S. A difference of 3.0 for the sum of propensities is associated with a higher proportion of the total amino acids associated with one of two states $\alpha$ or $\beta$ ($\alpha+\beta=$ 57%) than with the single state 'coil' (43%). This observation reinforces the choice of the current HCA alphabet VILF/MYW.

## Hydrophobic clusters are not ordinary patterns

**Hydrophobic cluster nomenclature.** A hydrophobic cluster always begins and ends with a hydrophobic amino acid coded 1. Other types of amino acids included within the cluster, with the exception of proline, are coded 0. The hydrophobic cluster is thus usually coded as a succession of 1 and 0 (e.g. 101011, fig. 4a). Together with the 2D shape, it can be further characterized by two additional codes. The P code

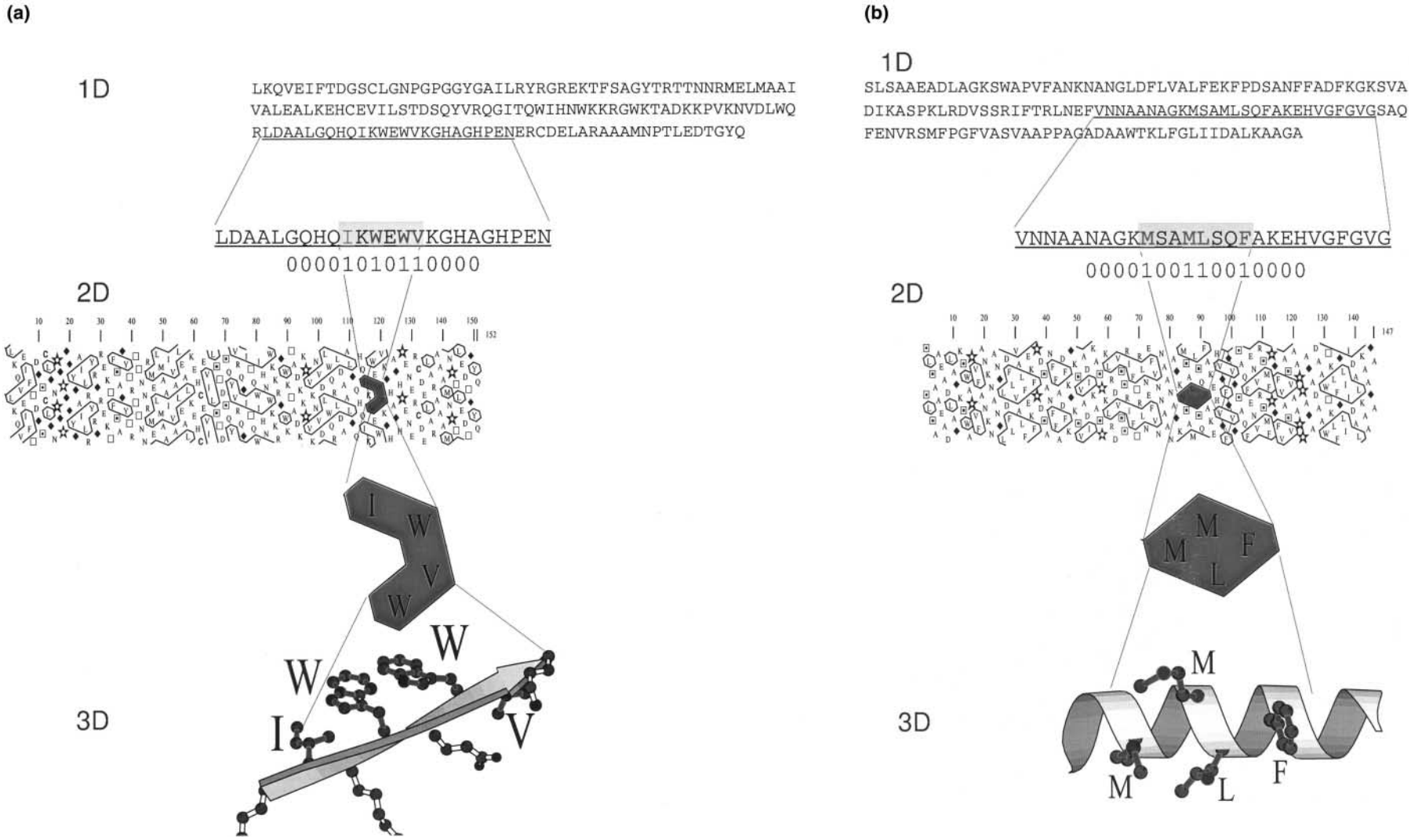


Figure 4. Illustration of some characteristics of two frequently encountered clusters. (*a*) The 101011 cluster (length six amino acids, four of which are hydrophobic). This cluster can be generated by sequences from 00001010110000 to P101011, its P code is 53 and its Q code is MMV. The ratio between observed occurrences (3419) and expected occurrences (3196) for random sequences (see text) is 1.07, with a Z-score of 4.4 $\sigma$. This cluster, mainly associated with $\beta$-strands, is thus slightly preferred to build proteins. An example of such a cluster is found in the ribonuclease structure. (*b*) The 10011001 cluster (length eight amino acids, four of which are hydrophobic). It can be generated by sequences from 0000100110010000 to P10011001P, its P code is 153 and its Q code is UVU. This cluster, mainly associated with $\alpha$-helices, is strongly preferred (2518 observed, 1260 expected, with a ratio of 2.0 and Z-score of $+35$ $\sigma$). An example of such a cluster is found in the myoglobin structure.

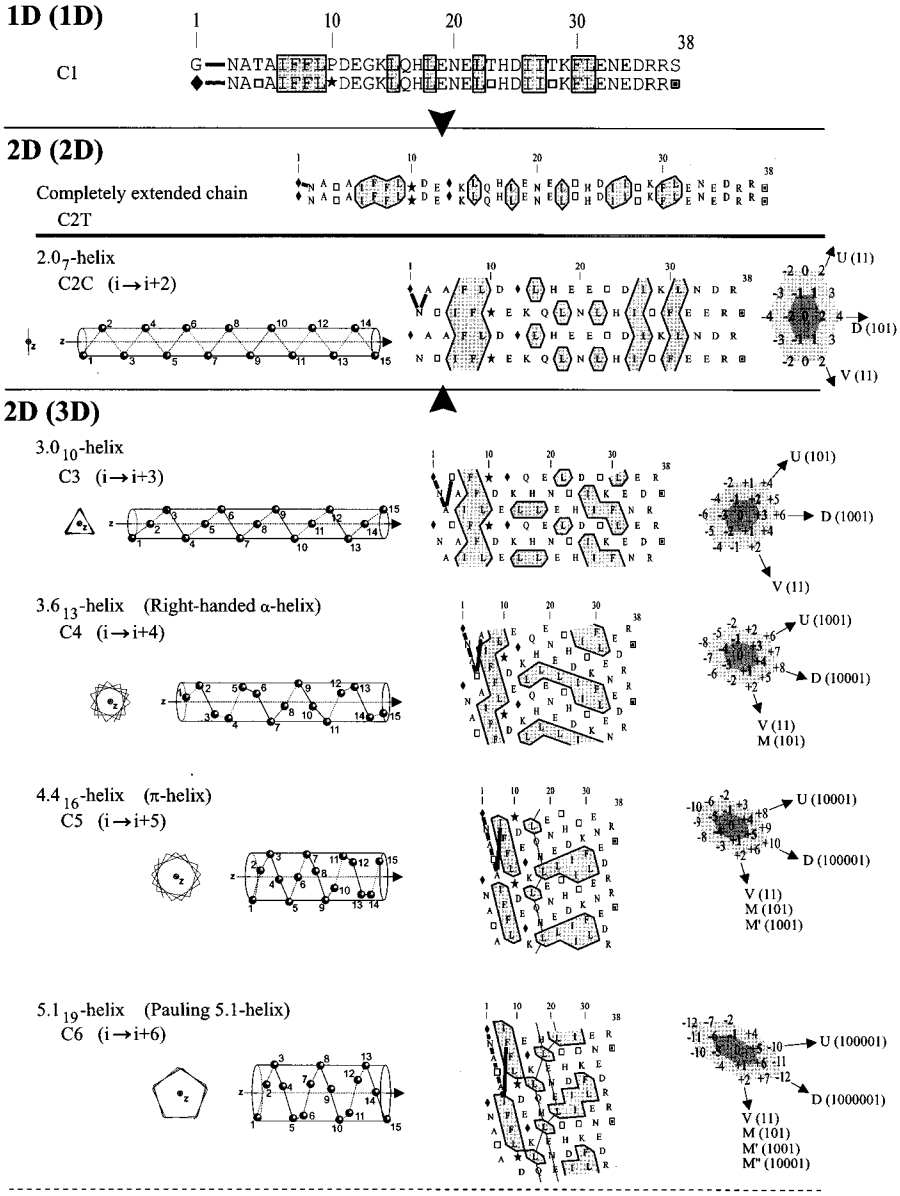


Figure 5. Same segment of human $\alpha$1-antitrypsin as in fig. 1 illustrating the formation of hydrophobic clusters for different helices with increasing connectivity distances (table 1). (C1) The simplest 'helix' C1, i.e. a straight line 1D immersed in a 1D space. It corresponds to the conventional representation of a sequence with a single first neighbour $i+1$ (G-N) and a connectivity distance of 1. (C2T) The C2T (T for Trans) configuration corresponds to the full extended polypeptide chain ($\phi$ and $\psi = 180°$), close to the $\beta$-strand conformation. First neighbour is also $i+1$. It corresponds to a 2D object immersed in a 2D space. (C2C) The C2C (C for Cis) configuration corresponds to the physically impossible helix possessing internal H-bonds between i and $i+2$ ($\phi$ and $\psi = 0°$). For this configuration, the helix reduces itself to a plane and can thus be attributed to the 2D plot/2D space. First neighbours on the 2D plot are $i+1$ and $i+2$. (C3) The $3.0_{10}$ helix which is actually the first 2D helical representation immersed in a 3D space with a connectivity distance of 3 ($i \rightarrow i+3$ H-bond). First neighbours on the 2D plot are $i+1$, $i+2$, $i+3$ (V, U and D, respectively) with no mosaic. (C4) The $\alpha$-helix with a connectivity distance of 4 ($i \rightarrow i+4$ H-bond). First neighbours on the 2D plot are $i+1$, $i+3$ and $i+4$, the mosaic corresponding to $i+2$. (C5) The $\pi$-helix with a connectivity distance of 5 ($i \rightarrow i+5$ H-bond). First neighbours on the 2D plot are $i+1$, $i+4$ and $i+5$, the mosaics corresponding to $i+2$ and $i+3$. (C6) The Pauling 5.1 helix [19] with a connectivity distance of 6 ($i \rightarrow i+6$ H-bond). First neighbours on the 2D plot are $i+1$, $i+5$ and $i+6$, the mosaics corresponding to $i+2$, $i+3$ and $i+4$. These constructions can be theoretically pursued. The bold line at the beginning of each plot indicates the increasing pitch of the corresponding helix. Water molecules could enter the growing central tunnel of the hypothetical helices larger than the C6 configuration (see fig. 6). Note: For current applications, the C4 HCA plot is shrunk along the horizontal direction to allow more sequence to be displayed in A4 paper format.

Table 1. Characteristics of polypeptidic helices on both sides of the $\alpha$ helix. Associated space is the original space in which the helix is drawn before being developed in a space of lower dimension. The pitch is the distance along the helix axis accomplished in one turn of helix. The connectivity distance, minimal number of nonhydrophobic amino acids that separate two successive HCA hydrophobic clusters (except for the case where a proline is present, see text), each cluster beginning and ending with a hydrophobic amino acid, is equal to the first neighbour distance in the D direction (fig. 4) and to the H-bond connectivity within the helices (see also fig. 5).

| Dimentionality and associated space | Helix | Pitch (Å) | Residues/ pitch | Theoretical helix diameter (Å) | Internal H-bond $C{=}O_{i} \cdots H{-}N_{i+n}$ | Residues superposed on the same generatrix of the helix |
|---|---|---|---|---|---|---|
| 0D (0D) | (one amino acid) | - | - | - | - | - |
| 1D (1D) | (the sequence) | - | - | - | - | - |
| 2D (2D) | completely extended chain (E) | 7.3 | 2.0 | - | - | - |
| | theoretical 2.0 helix | 4.6 | 2.0 | 3.0 | $i \rightarrow i+2$ | $i \rightarrow i+2$ |
| 2D (3D) | $3.0_{10}$ helix ($3_{10}$) | 6.0 | 3.0 | 3.7 | $i \rightarrow i+3$ | $i \rightarrow i+3$ |
| | right-handed $\alpha$-helix ($\alpha_R$) | 5.4 | 3.6 | 4.5 | $i \rightarrow i+4$ | $i \rightarrow i+18$ |
| | $\pi$-helix ($\pi$) | 5.1 | 4.4 | 5.5 | $i \rightarrow i+5$ | $i \rightarrow i+22$ |
| | pauling 5.1 helix | 5.0 | 5.1 | 6.3 | $i \rightarrow i+6$ | $i \rightarrow i+46$ |
| ... | ... | ... | ... | ... | ... | |

| Dimentionality and associated space | Helix | Connectivity distance | First neighbours | Estimated $C_\alpha - C_\alpha$ distances (Å) | Example | Helical configuration Cn |
|---|---|---|---|---|---|---|
| 0D (0D) | (one amino acid) | 0 | 0 | - | Gnataiff... | C0 |
| 1D (1D) | (the sequence) | 1 | $\pm(1)$ | - | GNataiff... | C1 |
| 2D (2D) | completely extended chain (E) | 1 | $\pm(1)$ | $\pm(3.8)$ | GNataiff... | C2T |
| | theoretical 2.0 helix | 2 | $\pm(1,2)$ | $\pm(3.8,4.6)$ | GNAtaiff... | C2C |
| 2D (3D) | $3.0_{10}$ helix ($3_{10}$) | 3 | $\pm(1,2,3)$ | $\pm(3.8,5.7,6.5)$ | GNATaiff... | C3 |
| | right-handed $\alpha$-helix ($\alpha_R$) | 4 | $\pm(1,3,4)$ | $\pm(3.8,5.6,6.3)$ | GNaTAiff... | C4 |
| | $\pi$-helix ($\pi$) | 5 | $\pm(1,4,5)$ | $\pm(3.8,4.9,6.4)$ | GNatAIff... | C5 |
| | pauling 5.1 helix | 6 | $\pm(1,5,6)$ | $\pm(3.8,5.7,6.3)$ | GNataIFf... | C6 |
| ... | ... | ... | ... | ... | ... | ... |

(first suggested by Manuel Peitsch, Glaxo Institute, Geneva, private communication) unequivocally codes the cluster with powers of 2 (for instance the 101011 cluster corresponds to a P code of 53 $((1 \times 1) + (2 \times 0) + (4 \times 1) + (8 \times 0) + (16 \times 1) + (32 \times 1))$. The Q code (Q for 'quark') results from the decomposition of any cluster into four basic clusters: **V** for 'vertical', 11, two consecutive hydrophobic amino acids; **M** for 'mosaic', 101; **U** for 'up', 1001 and **D** for 'down', 10001. (**V**/**M**), **U** and **D** constitute the three fundamental axes of the 2D $\alpha$-helical plot directed towards the first neighbours 1, 3 and 4 (fig. 1, [16]). Evidently, each of the V, D and U axes is a vectorial combination of the two others. The cluster 101011, which can be decomposed into 101 (**M**) concatenated with 101 (**M**) and 11 (**V**), then has the Q code **MMV**. The Q code also unambiguously characterizes a cluster.

In addition to the four fundamental clusters V, M, U and D described above, the singlet 1 (one isolated hydrophobic amino acid) is coded **S**.

**Difference between current patterns or sequence motifs and clusters.** The sequence fragment 101011, taken above as an example (fig. 4a), is found within a sequence database in many situations, either isolated from other hydrophobic amino acids or included in various other sequence fragments containing hydrophobic amino acids. In the HCA $\alpha$-helical plot, 101011 is a cluster only if it is separated from any other hydrophobic amino acid by at least four nonhydrophobic amino acids (0), except if at least one proline is present in positions $i-4$ to $i-1$ and $i+1$ to $i+4$. The number 4 corresponds to the connectivity distance of the $\alpha$-helix (see table 1). The cluster 101011 can be found in several environments, from the largest one 00001010110000 (no proline from $-4$ to $-1$ or from $+1$ to $+4$, on either side of the cluster) to the smallest one P101011P if two prolines directly flank the cluster.

Consequently, hydrophobic clusters constitute locally isolated entities far less numerous than common motifs or patterns of identical composition which do not obey connectivity rules (e.g. 101011 is found as a cluster 3419 times in a bank of 32,886 sequences and 45,746 times as a common motif). In contrast, they hold far more specific information, as already demonstrated [10] and observed in recent presentations of ordinary pattern properties [17, 18]. Indeed, the 2D $\alpha$-helical plot hydrophobic clusters best match the observed $\alpha$ or $\beta$ regular secondary structures of globular proteins. Statistically,

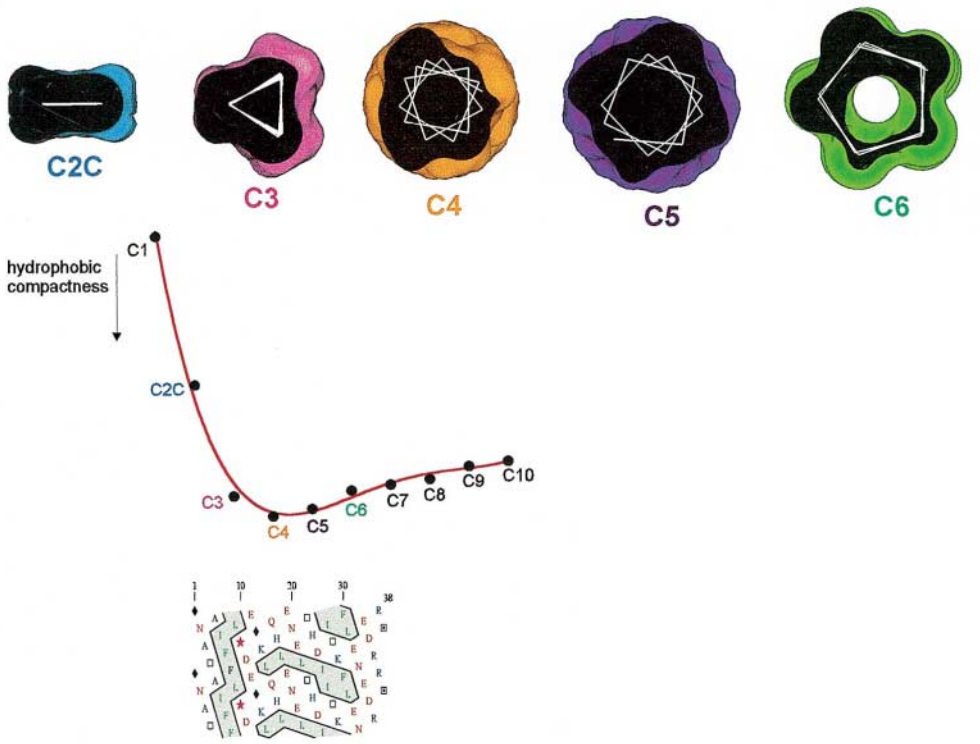


Figure 6. Illustration of the 2D hydrophobic compactness generated by different helical configurations from C1 to C10 (see table 1 and fig. 5). This compactness is measured by the total number of separate clusters (a mosaic cluster therefore counts for more than one cluster) calculated for each configuration in the nonredundant bank described in the legend of fig. 3. The sequence of fig. 5 taken as an example gives rise to 6, 6, 6, 4, 2, 3 and 4 compact clusters for the configurations C1, C2T, C2C, C3, C4, C5 and C6, respectively.

the gravity centres of hydrophobic clusters built with the hydrophobic alphabet VILFMWY exactly coincide with those of observed regular secondary structures, in perfect contrast, as expected, to clusters built with the typical loop alphabet PGDNS.

A frequently asked question about HCA is the apparent inadequacy of a helical plot to properly visualize and thus identify $\beta$-strands. In fact, this is actually not a problem, as the display of the sequence on a unique 2D $\alpha$-helical canvas provides a direct visualization of clusters associated with $\alpha$-helices (e.g. fig. 4b) and a transposed but codified visualization of clusters associated to $\beta$-strands (e.g. fig. 4a). Indeed, the correspondence between the clusters and secondary structure elements has been demonstrated to be statistically as efficient for all $\alpha$ as for all $\beta$ structures [10]. 2D helical plots different from the $\alpha$ one (as explained in table 1 and fig. 5) are not so efficient. From $\alpha$-helix (C4 helical configuration, H-bond between i and $i+4$, connectivity distance 4) towards greater pitches (smaller connectivity distances), one first encounters the $3_{10}$-helix (C3 helical configuration, H-bond between i and $i+3$, connectivity distance 3), and the overly constrained helix with an impossible H-bond between i and $i+2$ (connectivity distance 2). This structure corresponds to the C2C (C for Cis) helical configuration with $\phi$ and $\psi$ torsional angles equal to 0. The C2T (T for Trans) helical configuration ($\phi$ and $\psi$ equal to 180°) is the fully extended chain, conformationally very close to a $\beta$ strand. Finally one encounters the linear chain itself, i.e. the 1D sequence (connectivity distance 1). In the opposite direction of smaller pitches, one finds the C5 $\pi$-helix (H-bond between i and $i+5$, connectivity distance 5), the C6 Pauling 5.1 helix (H-bond between i and $i+6$, connectivity distance 6 [19]), and progressively more and more open helices (theoretically, for nonphysical helices, the opening and flattening of helices leads to a circle with an infinite radius and therefore a projection to a straight line perpendicular to the horizontal line of the 1D sequence).

The observation that it is the common $\alpha$-helix which shows the best concordance with experimental findings in exactly the same way for proteins of the different classes (all $\alpha$, all $\beta$, $\alpha/\beta$ or $\alpha+\beta$) [10] highlights an intriguing and compelling issue. This may perhaps be associated with the yet unknown folding rules, as nascent proteins may first favour local self-stable structures and particularly $\alpha$-helices. Interestingly, we observe in this context that it is the $\alpha$-helix (helical conformation C4 in table 1) which leads to a maximum

of hydrophobic 2D compactness (minimum of the number of separated hydrophobic clusters). In this respect, C4 is two times better than C1 (fig. 6).

**Mosaic clusters.** Many clusters display an isolated hydrophobic amino acid (or group of amino acids) in position $i \pm 2$ which does not correspond to first neighbours in the plot (see figs 1 and 5). Such clusters which contain a regular alternation of hydrophobic and non-hydrophobic amino acids are named 'mosaic clusters' and visualized using connecting lines (e.g. the first cluster of fig. 2). These connecting lines help to identify these special clusters which can therefore be formally separated from the true 'compact' clusters as defined by Bourat and coworkers [16]. Small mosaic clusters (e.g. 101, 10101) are often associated with $\beta$-strands and especially with surface ones.

**'Identity clusters'.** When comparing sequences through HCA (see below), one often encounters, relative to anchor references made of alignable hydrophobic clusters, several patches of amino acids which are chemically conserved. These 'identity clusters' can be visualized as illustrated in fig. 13 (pink shaded). These occurrence of such 'identity clusters', along with the hydrophobic clusters, are very useful in recognizing distant evolutionary relationships between proteins.

**Transmembrane segments and membrane proteins.** HCA plots are also useful for analysing transmembrane segments and membrane proteins. One should note that the hydrophobicity of each segment has to be appreciated relative to the type of protein considered. One isolated anchor transmembrane segment within a protein is often totally hydrophobic (V, I, L, F, M, Y, W and mainly P, C, A, S, T). Several associated transmembrane segments often share a few hydrophilic amino acids which together may constitute internal polar active sites (for example the receptors which possess seven transmembrane segments). Loops between transmembrane segments can be analysed taking into account that they often do not constitute classical globular domains, except if they are of sufficient length.

**Cluster lengths and secondary structures.** HCA hydrophobic clusters are statistically centred on internal faces of regular secondary structures $\alpha$ and $\beta$. However the length of the clusters may be larger than those of single secondary structures. Fig. 1 illustrates such a case, with one cluster associated with two successive helices. Indeed, one or several hydrophobic amino acids (V, I, L, F, M, Y, W) may link two clusters (rarely more) giving rise to a larger one. It is often possible to detect such concatenations by comparison of related sequences and/or by analysis of the shape and composition of the clusters. Amino acids such as Y, G, N, D and S between two compact sub-clusters, and associated with a change in the main cluster direction, are often indicative of artificial concatenation. The automatic processing of clusters, from sequences only, in order to fit their length better (by decreasing or increasing it) to the length of regular secondary structures in a context-dependent manner, has given encouraging results (E. Thoreau, unpublished results).

## Protein sequences are not random distributions of amino acids

Many properties of hydrophobic clusters can be illustrated through statistical studies of sequences and 3D

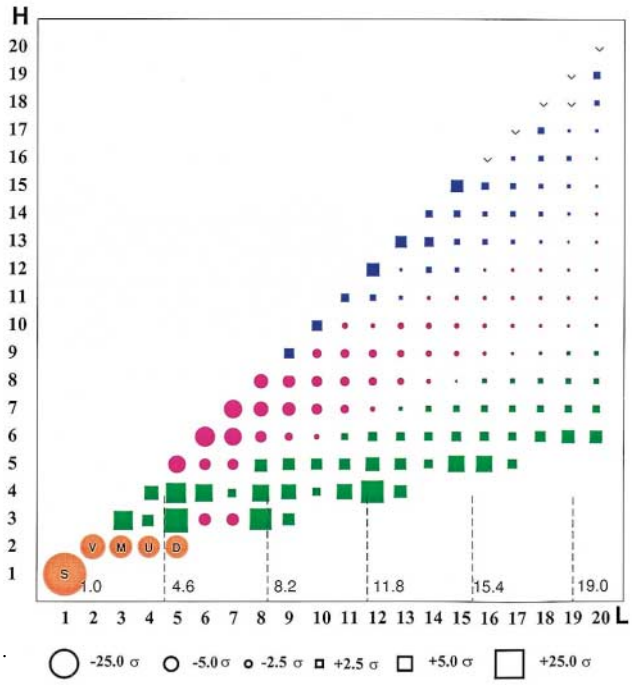


Figure 7. Global map of protein sequence properties. Plot of the mean behaviour of hydrophobic clusters (built with the VILFMYW alphabet) belonging to each (L,H) class. L is the amino acid length of the cluster, H is the number of hydrophobic amino acids. Squares indicate preferred positions, the surface being proportional to the mean Z-score. Circles indicate relatively discarded positions. The diagonal is occupied by clusters only composed of hydrophobic amino acids 1, 11, 111, etc... (one kind per position). The other occupied positions have a variable number of isotypic cluster kinds (same L, H values). For example in the position $L = 8$, $H = 4$, there are 12 possible isotypic clusters, one of which corresponds to the losange 10011001 shown in fig. 3b. The number of different kinds of possible clusters increases very rapidly with L, H values (e.g. 216 for $L = 12$, $H = 7$). The map shown here contains about $10^6$ clusters. Several striking features can be highlighted: the poor representation of the small basic clusters S, V, M, U, D (bottom left corner), a wide favourable area (green squares) reinforced at the $\alpha$-helix periodicity (4.6, 8.2, 11.8, 15.4, 19.0 amino acids) and corresponding to the preferred bricks of globular domains, an unfavourable area (purple circles) above and another favourable area (blue squares) which corresponds to stretches of transmembrane segments (e.g. 12 consecutive hydrophobic amino acids). The reinforcement of cluster properties with the $\alpha$-helix periodicity and a clear interruption after $L = 20$ are also observed for several other properties (data not shown).

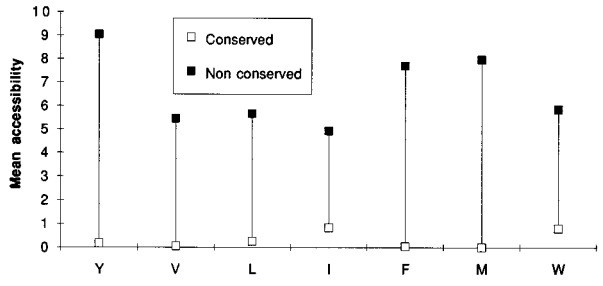


Figure 8. Mean accessibilities (in $Å^2$) of conserved hydrophobic positions (positions where only hydrophobic amino acids are found) and nonconserved ones in the tryptase family (PDB identifiers: 1AAO, 1ARB, 1CHG, 1ELD, 1ESA, 1GCT, 1HCG, 1HF1, 1MCT, 1NTP, 1P12, 1PCU, 1PFA, 1PPB, 1PPF, 1SGT, 1TON, 2ALP, 2CP1, 3RP2, 4PTP, 4SGA, 8GCH). All the sequences of the family have less than 50% identity with all the others. There are seven absolutely conserved hydrophobic positions; the mean length of the sequences of the family is 232.

structures of proteins. These studies led to the characterization of unexpected features associated with the mechanisms of globular domain stability. A few of these features are described below.

One of the more readily obtained properties concerns the difference between the occurrence of the different kinds of clusters and their expected occurrence in random sequences with an identical chemical composition. In order to confirm such observations, a large bank of protein sequences has been constructed based on the Genpep databank (1994 fall issue, 155,144 entries). To avoid possible statistical bias due to the large redundancy of such a bank, a mainly nonredundant sequence bank has been extracted containing a total of 32,886 entries (11,078,281 amino acids). The observed occurrence of all kinds of hydrophobic clusters with lengths ranging from 1 to 20 amino acids (more in many cases) has been calculated and compared to the expected occurrences calculated on the basis of random positions of amino acids assuming chemical conservation of each sequence (100 randomizations per sequence). The difference between the observed and expected occurrences is quantified by their ratio and by their Z-scores (expressed in standard deviation units).

Nature clearly has a preference for many clusters (Z-scores up to $+35\ \sigma$) and has preferentially discarded others (Z-scores up to $-65\ \sigma$). Ratios greater than 2.0 and less than 0.5 are frequent. Figs 4a and 4b illustrate such characteristics for two clusters, one (10011001) mainly associated with $\alpha$-helices, the other (1010011) preferentially associated with $\beta$-structures. Each cluster is a structural brick for which hydrophobic side chains together constitute 'scratch' areas which 'zip' each other to form the hydrophobic core of globular domains.

Fig. 7 summarizes the mean behaviour of all hydrophobic clusters built with the C4 helical configuration, the current alphabet VILFMYW and possessing lengths ranging from 1 to 20 amino acids. It represents a global insight into how nature has selected structural bricks to build proteins in a water environment.

Such maps have been calculated for clusters constructed with other alphabets in order to highlight the role of each amino acid. For example, the replacement of only one amino acid of the conventional alphabet VILFMYW (L by E) completely destroys the patterns of fig. 7, while replacement of Y or W by C is almost indifferent and even locally reinforces the differences between preferred and discarded positions.

## Hydrophobic clusters are tightly associated with secondary structures and their structural 3D properties

Five years ago it was observed and demonstrated statistically [10] that the 2D hydrophobic clusters displayed on the $\alpha$-helical net coincide mainly with regular secondary structures $\alpha$ or $\beta$. Moreover, many types of clusters show marked preference to be associated with one of these two states. In contrast, small clusters possessing a low number of hydrophobic amino acids, such as the basic clusters **V**(11), **U**(1001), **D**(10001) are often associated with coil regions (loops). As an example, the 10011001 cluster illustrated in fig. 4b exhibits a strong preference for the $\alpha$ state (79% for its central part), in contrast to its occurrence in $\beta$ (10%) and coil (11%) structures. This kind of data can now be derived from the rapidly growing number of 3D structures of proteins, even if many of them are clearly redundant. This information opens new perspectives for the comparison of sequences and the prediction of their associated 3D structures.

The frequency with which all current clusters in the sequence banks, and a great many of them in the structure databases, occur is high enough to allow statistical observations for each amino acid in each position of a cluster as well as for its proximal neighbours (between $i-4$ to $j+4$).

## HCA can distinguish several populations of hydrophobic amino acids within globular domains

The HCA alignment of structurally related but divergent sequences makes it possible to delineate precisely the positions which are always or often occupied by hydrophobic amino acids. These segregate into two main populations, clearly exhibiting different properties concerning, in particular, their solvent accessibility as well as their geometric arrangement around the mean structure. The conserved hydrophobic positions (often chemically different) are much more buried than the nonconserved ones and their side chains are signifi-

Table 2. Selected HCA applications performed by our groups or in collaboration.

Glycosyl hydrolases/Glycosyl transferases

- **Henrissat B., Clayssens M., Tomme P., Lemesle L. and Mornon J.-P.** (1989) Cellulase families revealed by Hydrophobic Cluster Analysis. Gene **81:** 83–95. The first paper related to the sequence analysis of glycosyl hydrolases. Prediction of the catalytic amino acids verified by directed mutagenesis by Py et al. (1991) Prot. Eng. **4:** 325–333.
- **Henrissat B., Callebaut I., Fabrega S., Lehn P. and Mornon J.-P.** (1995) Conserved catalytic machinery and prediction of a common fold for several families of glycosyl hydrolases. Proc. Natl Acad. Sci. USA **92:** 7090–7094. Classification of several families of glycosyl-hydrolases in the clan 'GH-A' and prediction of a similar fold and catalytic machinery for several other families, for which the catalytic amino acids were unknown. These predictions were later assessed by the experimental three-dimensional structure of several of the described enzymes.
- **Saxena I., Brown R. M. Jr., Fèvre M., Geremia R. A. and Henrissat B.** (1995) Multiple domain architecture of $\beta$-glycosyl transferases. Implications for mechanism of action. J. Bacteriol. **177:** 1419–1424. Distinction of two groups among glycosyl transferases forming $\beta$-glycosylic bonds from a nucleotide-sugar unit: those having two domains (A and B) or only one (A). The latter can only add one sugar whereas the former are able to polymerize. This observation has led to the proposal that polysaccharide biosynthesis by enzymes which possess the two domains probably occurs by simultaneous addition of two nucleotide-sugars.
- **Bolam D. N., Hughes N., Virden R., Lakey J. H., Hazlewood G. P., Henrissat B. and Gilbert H. J.** (1996) Mannanase A from Pseudomonas fluorescens subsp. cellulosa is a glycosyl hydrolase in which E212 and E320 are the putative catalytic residues. Biochemistry **35:** 16195–16204. Assignment of the family 26 of glycosyl hydrolases to the clan GH-A. Identification of the two catalytic residues.
- **Fernandes M. J. G., Leclerc D., Henrissat B., Vorgias C. E., Gravel R., Hechtman P. and Kaplan F.** (1997) Identification of candidate active site residues in lysosomal $\beta$-hexosaminidase. J. Biol. Chem. **272:** 814–820. Identification of two catalytic residues of a lysosomal enzyme (prediction through HCA and directed mutagenesis experiments).
- **Durand P., Lehn P., Callebaut I., Fabrega S., Henrissat B. and Mornon J.-P.** (1997) Active site motifs of lysosomal acid hydrolases: invariant features of Clan GH-A glycosyl hydrolases deduced from Hydrophobic Cluster Analysis. Glycobiology **7:** 277–284. Sequence and structural characterization of the Clan GH-A. Implications for several important lysosomal enzymes responsible for lysosomal storage diseases.

Cytokine/Growth hormone/Prolactin receptor family

- **Gaboriaud C., Uzé G., Lutfalla G. and Mogensen K.** (1990) Hydrophobic cluster analysis reveals duplication in the external structure of human alpha-interferon receptor and homology with gamma-interferon receptor external domain. FEBS Lett. **269:** 1–3.
- **Thoreau E., Petridou B., Kelly P. A., Djiane J. and Mornon J.-P.** (1991) Structural symmetry of the extracellular domain of the Cytokine/Growth hormone/Prolactin receptor family and Interferon receptors revealed by Hydrophobic Cluster Analysis. FEBS Lett. **282:** 26–31. Identification of a duplicated module, related to the immunoglobulin fold, in the extracellular part of the Cytokine/Growth hormone/Prolactin receptors. This prediction was independently proposed by Bazan J. F. (1990) Proc. Natl Acad. Sci. USA **87:** 6934–6938 and verified by the X-ray structure of the growth hormone receptor (De Vos et al. (1992) Science **255:** 306–312).
- **Vigon I., Mornon J.-P., Cocault L., Mitjavila M. T., Tambourin P., Gisselbrecht S. and Souyri M.** (1992) Molecular cloning and characterization of h-mpl, the human homolog of the v-mpl oncogene: identification of a new member of the hematopoietic growth factor receptor superfamily. Proc. Natl Acad. Sci. USA **89:** 5640–5644.

FKBP family

- **Callebaut I., Renoir J. M., Lebeau M.-C., Massol N., Burny A., Baulieu E. E. and Mornon J.-P.** (1992) An immunophilin that binds Mr 90,000 heat shock protein: main structural features of a mammalian p59 protein. Proc. Natl Acad. Sci. USA **89:** 6270–6274. Characterization of HBI, the first identified FKBP member with two FKBP modules associated with nontransformed steroid receptor complexes.
- **Callebaut I. and Mornon J.-P.** (1995) Trigger factor, one of the Escherichia coli chaperone proteins, is an original member of the FKBP family. FEBS Lett. **374:** 211–215. Identification of the trigger factor as a member of the FKBP family. Characterization of its active site. The PPIase activity of the trigger factor as well as its importance in protein folding was later demonstrated by several groups (see text).
- **Blecher O., Erel N., Callebaut I., Aviezer K. and Breiman A.** (1996) A novel plant peptidyl-prolyl cis-trans isomerase (PPIase): cDNA cloning, structural analysis, enzymatic activity and expression. Plant Molec. Biol. **32:** 493–504. Characterization of the first plant member of the FKBP family, sharing similarities with HBI but having three FKBP modules instead of two.

Cell cycle control – breast cancer

- **Callebaut I. and Mornon J.-P.** (1997) From BRCA1 to RAP1: a widespread BRCT module closely associated to DNA repair. FEBS Lett. **400:** 25–30. Identification of a large superfamily of proteins containing one or several modules of a 'BRCT' domain. These include the breast cancer susceptibility antigen BRCA1 protein, rad4, rad9, xrcc1, ect2, the DNA ligases III and IV, RAP1 and TdT. Most of them are involved in cell cycle regulation and response to DNA damage. This prediction was simultaneously published by Bork and coworkers (FASEB J. (1997) 11, 68–76).

Viruses and related topics

- **Callebaut I., Portetelle D., Burny A. and Mornon J.-P.** (1994) Identification of functional sites on bovine leukemia virus envelope glycoproteins using structural and immunological data. Eur. J. Biochem. **222:** 405–414. Modeling of the envelope glycoproteins of an oncovirus.
- **Callebaut I., Tasso A., Brasseur R., Burny A., Portetelle D. and Mornon J.-P.** (1994) Common prevalence of alanine and glycine in mobile reactive centre loops of serpins and in viral fusion peptides. Do prions possess a fusogenic peptide? J. Comp. Aided Molec. Design **8:** 175–191. Characterization of the fusion peptides of retroviruses, orthomyxoviruses and paramyxoviruses and role in their conformational mobility (see also fig. 9).
- **Bénit L., de Parseval N., Casella J.-F., Callebaut I., Cordonnier A. and Heidmann T.** (1997) Cloning of a new murine endogenous retrovirus, MuERV-L, with a strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J. Virol. **71:** 5652–5657. Identification of the 'MHR' signature in the gag coding sequence of a murine endogenous retrovirus closely related to the mouse retrovirus restriction gene Fv1. This latter gene, conferring resistance to murine leukemia viruses in the early stages of infection, was recently isolated (see Best et al., Nature (1996) **382:** 826–829 and Goff S. P., Cell (1996) **86:** 691–693). At the sequence level, assumptions on the gag function of the Fv1 gene were only based on its similarity to the HERV-L family of human endogenous retroviruses and on its position in the element. Here we show that the Fv1 gene, as well as the gag portion of HERV-L and MERV-L, harbours the signature of the retroviral gags centred around the 'major homology region', despite no detectable similarity by conventional methods.

Table 2 (continued).

Other topics

- **Schoentgen F., Seddiqi N., Bucquoy S., Jollès P., Lemesle-Varloot L., Provost K. and Mornon, J.-P.** (1992) Main structural and functional features of the basic cytosolic bovine 21 Kda protein delineated through Hydrophobic Cluster Analysis and molecular modelling. Prot. Eng. **5:** 295–303.
- **Labesse G., Kotoujansky A., Chomilier J. and Mornon J.-P.** (1993) The VirC1 protein of Agrobacterium tumefaciens belongs to a family of ATP-binding proteins involved in active plasmid partition. Protein Seq. Data Anal. **5:** 345–348.
- **Labesse G., Vidal-Cros A., Chomilier J., Gaudry M. and Mornon J.-P.** (1994) Structural comparisons lead to the definition of a new superfamily of NAD(P)(H)-oxydoreductases: the single-domain Reductases/Epimerases/Dehydrogenases; the RED family. Biochem. J. **304:** 95–99.
- **Callebaut I. and Mornon J.-P.** (1997) The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in Drosophila melanogaster development. Biochem. J. **321:** 125–132. Identification of a repeated module related to the OB-fold in the human EBNA-2 coactivator p100. Identification of a novel domain largely repeated in the Drosophila melanogaster tudor protein. After the publication of this reference, these results were independently reported by Ponting (1997) Protein Science **6:** 459–463 and Trends Biochem. Sci. **22:** 51–52.
- **Ye Q., Callebaut I., Pezhman A., Courvalin J.-C. and Worman H. J.** (1997) Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. J. Biol. Chem. **272:** 14983–14989. Delineation of interacting domains within the HP1 and LBR proteins.

cantly less exposed. The small number of fully or almost fully conserved positions constitutes a strong determinant of each fold (fig. 8).

## HCA as a tool to compare protein sequences within and below the twilight zone

Many HCA applications have been successfully performed during these last few years, leading to the deciphering of key structural and functional features of a large number of proteins. Some striking examples, performed by our groups or in collaboration, are briefly described in table 2. Here, we would like to give a few typical examples to illustrate the effectiveness of the method in detecting similar folds or similar motifs between sequences showing very limited sequence relatedness (typically below the twilight zone of 25–30% sequence identity).

The method is already valuable in the study of the protein sequence in itself, before any comparison step. The first 'crude' examination of a sequence texture through its HCA diagram generally allows its 'preprocessing' by defining the domain organization of the protein, as examplified in fig. 9 for prion. Globular domains, which are characterized by a typical thick distribution of hydrophobic clusters, can easily be recognized and delineated (see region (4) in fig. 9). They are often separated from other domains by hydrophilic or weakly hydrophobic regions of variable length ('hinges', see fig. 10 with one member of the BRCT superfamily, see below). Similarity searches in data banks can then be restricted to targeted regions, thereby avoiding regions of compositional bias such as hinges, membrane-spanning and coiled-coil regions.

The detection of internal repeated domains, as illustrated in fig. 10 and hereafter with the example of cytokine receptors (fig. 11), is consequently one of the most direct results which can be deduced from an overall analysis of a protein sequence through HCA.

The extracellular domain of the large superfamily which includes several cytokine receptors as well as the growth hormone and prolactin receptors is composed of one or several copies of a domain of approximately 200 amino acids. This 'basic' domain was in fact found to be composed of a repeat of two structurally similar units of 100 amino acids [27]. Despite a relatively good cluster conservation between the two subdomains in the family, sequence identities (identical residues shaded in fig. 11) are very low and often below 10%. Most of the clusters have typical features of $\beta$-strands (vertical and mosaic shapes), suggesting an all-$\beta$ fold, probably of the Ig-like type as suggested by the occurrence of several members of this folding family in the FASTA and BLAST outputs, although only through nonsignificant hits. The retrieval of a well-conserved mosaic motif of the Ig-like superfamily, the Y-{VILMFYW}-C motif of strand F (where curly brackets indicate the amino acids which are forbidden in a given position), with the C replaced here by any hydrophobic residue, strengthens this prediction, subsequently confirmed by the experimental structure of the growth hormone receptor [28].

Given the complexity of the various and unpredictible issues that have to be addressed at low levels of identity, fully automatic methods for database scanning based on HCA principles, although desirable, are not yet feasible. With the exception of the detection of internal repeats (see the examples of the cytokine receptors and chromo superfamilies in table 2) or functional clues that can orientate the search (glycosyl hydrolases in table 2), HCA is thus routinely used in combination with 1D methods (fig. 12). The results of searches with the current BLAST [5, 29] and FASTA [6] programs are starting points for the HCA analysis. The particular efficiency of HCA stems from its ability to distinguish authentic 3D relationships from statistically insignificant BLAST and FASTA hits (for example with P

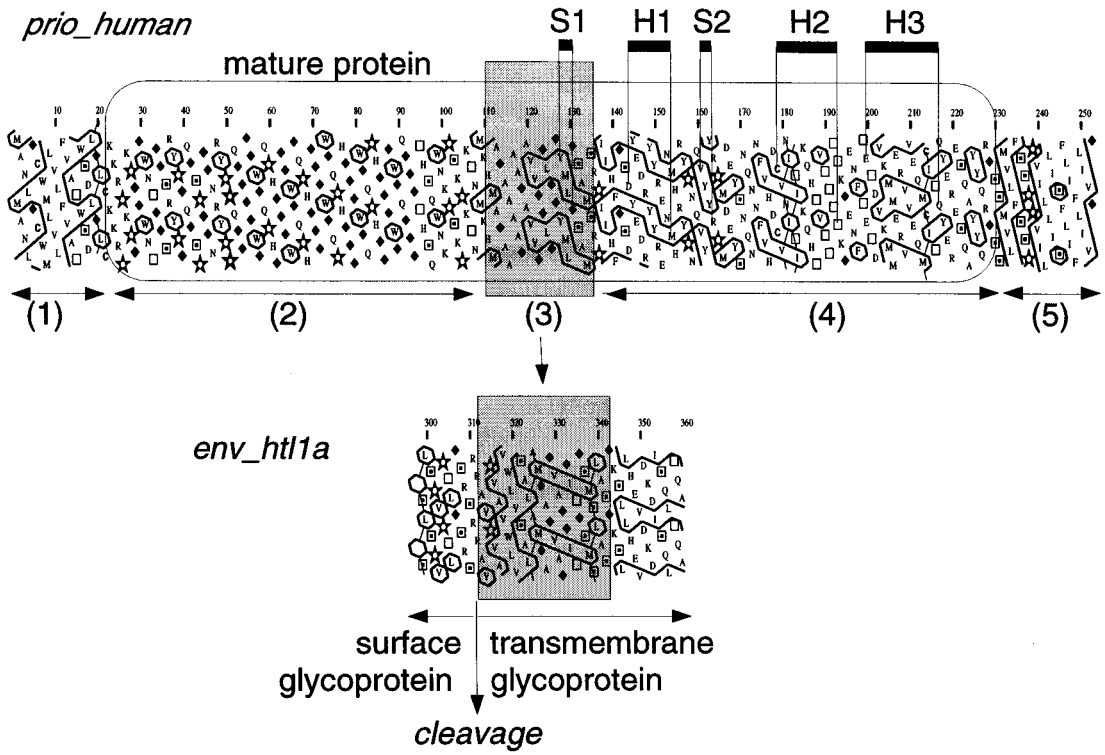


Figure 9. One example of rapid and visual delineation of domains through HCA. The sequence of the human prion (PrP), as found in the Swissprot databank (prio_human), can rapidly be divided into five distinct domains: (1) a largely hydrophobic region corresponding to the signal peptide, (2) a nonglobular region characterized by many short repeats, (3) a nonhydrophilic region containing hydrophobic amino acids as well as short residues such as alanine and glycine, (4) a globular region of approximately 100 amino acids, and (5) a hydrophobic peptide typically corresponding to a membrane-spanning domain. The two terminal regions (1) and (5) are removed in the mature protein (boxed). Prion proteins are thought to exist in two different conformations: the 'benign' $PrP^{c}$ form and the 'infectious scrapie' form, $PrP^{Sc}$. Transition from $PrP^{c}$ to $PrP^{Sc}$ has been shown to be linked to an increase in the $\beta$-sheet content of the protein [20]. Several years ago, we noticed that the global composition of the third region is reminiscent of that of fusion peptides from viral envelope glycoproteins (the fusion peptide of HTLV-I is shown below), suggesting that this region, like fusion peptides, may also be structurally highly mobile [21]. Both peptides are hydrophobic and especially rich in small amino acids such as glycine and alanine. This region could therefore play an essential role in the conformational change observed in the $PrP^{c}$ to $PrP^{Sc}$ transition. These hypotheses were subsequently supported by the experimental demonstration of neurotoxicity associated with a fragment corresponding to this region [22] and of its conformational polymorphism [23, 24]. The basic structure of the mouse PrP globular domain encompassing the second half of region (3) as well as the entire region (4) and containing most of the point-mutation sites that have been linked, in human PrP, to the occurrence of familial prion diseases, has recently been solved [18], showing a two-stranded anti-parallel $\beta$-sheet (S1 and S2) and three $\alpha$-helices (H1 to H3; secondary structures are reported above the HCA plot). It has been speculated that the short $\beta$-sheet, including the second part of the third PrP region (3), might be a 'nucleation site' for the conformational transition [25], thereby reinforcing our hypothesis of a crucial role for this segment which otherwise possesses a high glycine content, unusual in classical $\beta$-strands. Other regions can be highlighted from the HCA plot, such as that corresponding to the helix H2 (amino acids 179 to 193) which also possesses an amino acid composition and cluster shape typical of $\beta$-strands and not of an $\alpha$-helix (see fig. 3), and which therefore could also be involved in the conformational transition.

values = 1 in the BLAST outputs), by putting the observed similarities in the context of the 2D structure. It is also common that only a part of the similarity region is detected by the BLAST and FASTA searches, usually focusing on the most conserved site(s) of a domain (catalytic site for example). In this context, HCA can often extend the similarity region beyond the regions retained by these 1D screening methods, as shown in fig. 13. This extension often strengthens a possible similarity which otherwise would remain indistinguishable from spurious ones. Then, using an iterative strategy like that used to assemble a jigsaw puzzle (fig. 12), the retained hits are again searched against databases or aligned to derive a characteristic profile which can also be searched. In this respect, both methods, i.e. those producing pairwise alignments and those working with motifs or profiles (reviewed in [8]), are often complementary, especially in detecting highly divergent domains. The success of the searches, using both strategies, often depends on the divergence of the sequences chosen to constitute a 'searching' set, as greater divergence eliminates redundant information. In particular, in some sequences it is useful to 'follow' and retrieve the main features of clusters which appear to be

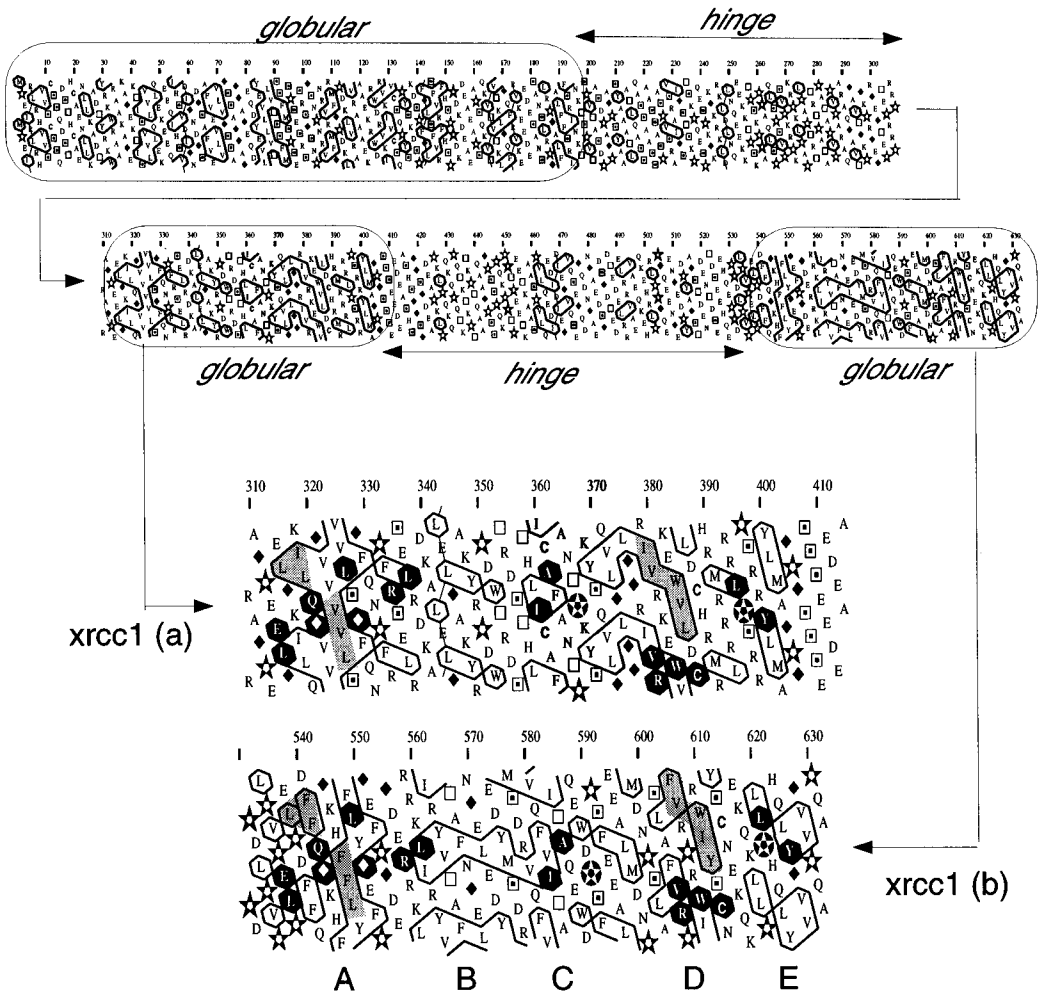


Figure 10. The human xrcc1 protein (SwissProt xrcc_human) is composed of three globular domains clearly separated by two hinge regions. The last two domains, sharing 19% sequence identity, are structurally similar and correspond to BRCT modules ([26], see also fig. 11). Five motifs, designated A to E, are found conserved in the whole family. The conservation of motifs A and D is clearly visible here.

characteristic of the family. Evolution scatters these clusters within the family and thus considerably helps the alignment process.

One of the limiting aspects of the analysis of a sequence showing 'no obvious similarity' with the sequences contained in the databases concerns the detailed analysis of the BLAST and/or FASTA outputs, which are often difficult to assimilate in the absence of any 'functional' clue. Fortunately, several programs designed to handle such output files now exist, helping to choose the candidates to be analysed by HCA. For example, the VISUALBLAST and VISUALFASTA programs [30] developed in our laboratory provide an informative and interactive output of the BLAST and FASTA results, respectively, with direct links to HCA. MULBLAST [31], another tool recently designed in our laboratory, extracts multiple alignments from BLAST output files, thereby helping to identify conserved regions and allowing new protein motifs to be described. The deduced gapped multiple alignments can then be submitted to profile searches. Such searches can also be performed with multiple alignments blocks [32]. Other tools such as FTHOM [33], BLA [34] and BEAUTY [35] use the information contained in the SWISSPROT and PROSITE databases to identify already known domains or patterns from the BLAST outputs.

Below 25–30% sequence identity, when 'true' relationships are not easily distinguished from background noise, two different levels of analysis can be performed:

– Typically, down to 15% identity and often below, precise sequence alignments can be used for valuable 3D modelling as well as, in the best cases, molecular replacement for crystallographic structure determinations [36]. The predictions which have been tested by

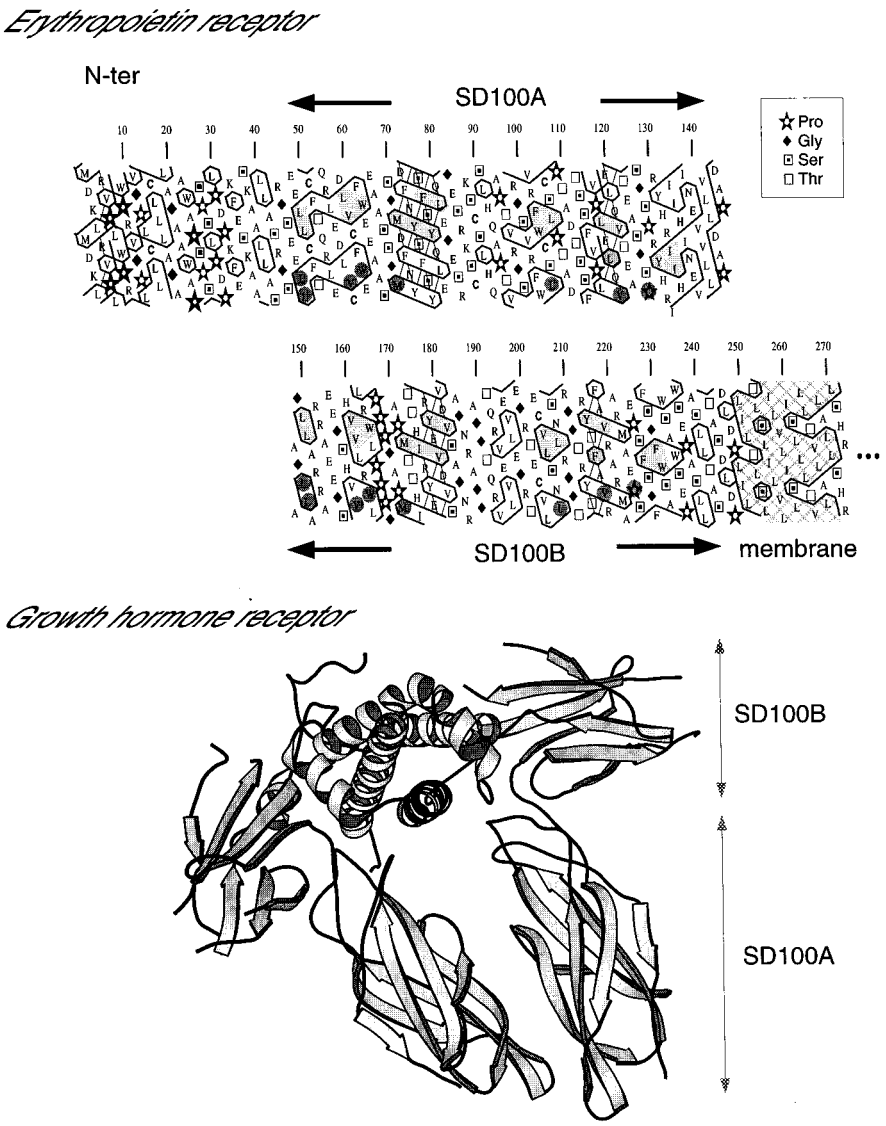


Figure 11. The extracellular 200 amino acid module of cytokine receptors, as illustrated here with the erythropoietin receptor, was found to consist of a duplicated module of ~100 amino acids whose fold has been predicted to be of the immunoglobulin-type. Cluster similarities are shown in light shading, sequence identities (below 10%) in dark shading. This prediction has been verified by the experimental structure of the growth hormone receptor [28], represented here in its dimeric form and complexed with the hormone.

directed mutagenesis or experimental structure determination have been verified.

– Below 15% sequence identity, accurate predictions are often limited to the most conserved regions such as catalytic sites, as exemplified by the analysis of glycosyl hydrolases families (see table 2). In a few cases, the general features of the fold adopted by several families of glycosyl hydrolases have been successfully predicted from less than 5% sequence identity. The difficulty of producing a precise alignment below 15% sequence identity depends on the nature of fold, as other domains, such as those of the cytokine receptors and of the BRCT family (see hereafter), can be aligned with precision despite very low levels of sequence identity.

Only two predictions, performed at very low sequence identity and described further below, have not been verified experimentally.

Frequently, the amino acid sequence deduced from a newly sequenced gene is reported to share 'no obvious similarity to any other gene found in sequence databases'. Consequently, several crucial genes such as susceptibility antigens for important diseases remain orphaned of function(s). On the other hand, the rapid emergence of complete genomes has left a considerable number of unidentified genes, many of which probably encode multidomain proteins. The following three recent examples demonstrate that further examination

through adapted methods such as HCA can often help to suggest a function for these unclassified proteins and to highlight presumably important motifs.

The first example concerns a human p100 protein which has been described as a coactivator of gene expression induced by the Epstein-Barr virus nuclear antigen 2 (EBNA 2) [37]. This protein is also able to bind DNA. We have shown that p100 consists of a repeat of five similar domains, the fifth being considerably modified relative to the other four [38] (fig. 14). The fold of the p100 repeated module can be related to that of the staphylococcal nuclease (thermonuclease-SN), whose first subdomain (not shaded in fig. 14) belongs to the large OB fold (Oligonucleotide-Oligosaccharide Binding) superfamily [39]. The compatibility of the 2D signature of the p100 SN-like domains with that of SN, sharing between 19.5 and 20.8% sequence identity, was checked with HCA. Although the SN fold is well conserved, the SN catalytic amino acids are missing, suggesting that the p100 OB fold could only serve to bind DNA without catalytic activity, as with many other OB folds.

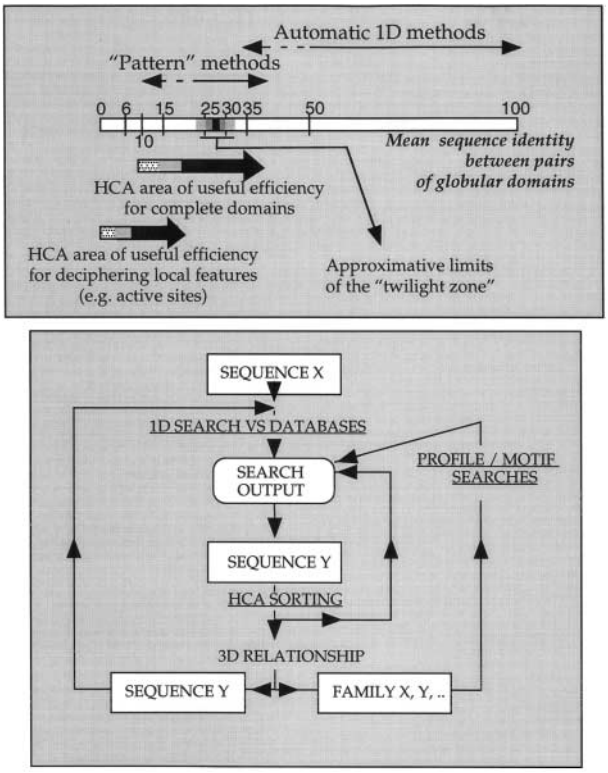


Figure 12. Strategy for database searching using HCA. HCA is very useful within and below the so-called 'twilight zone', as illustrated in the selected applications of table 2. With more than 15% sequence identity on a sufficient length (typically the length of a globular domain), no false positives have yet been observed. Below this level of sequence identity, very few false positives have been observed (see text) and many accurate predictions have been made (e.g. [27]). In this area, HCA is particularly efficient in identifying locally conserved features such as catalytic amino acids (e.g. glycosyl hydrolases in table 2) due to its ability to anchor alignments relative to 2D features independently of insertions and deletions and background noise. Note that the mean sequence identity of pairwise alignments within a family of structurally related domains covers a large area of individual values. It is therefore essential to use the maximum number of potentially related sequences to overcome distance between very remote members.

This first subdomain of the p100 fifth repeat has been replaced by a module found in multiple repeat copies in the *Drosophila melanogaster* tudor protein and therefore named the 'tudor' domain [38]. The tudor module belongs to an as yet uncharacterized and highly divergent family (below 10% sequence identity) [38, 40].

The trigger factor is one example of a protein whose mechanism of action has long remained obscure. Trigger factor is an abundant soluble protein originally discovered in *Escherichia coli* in 1987 [41–44]. It was described as a chaperone protein which forms soluble complexes with the precursor of outer membrane protein A (OmpA) and assists in the maintenance of translocation competence. In 1995, we showed that the central domain of this protein of 432 amino acids belongs to the FKBP (FK506-Binding Protein) family and consequently predicted that it could function as a rotamase [45] (fig. 15).

The highest sequence identities observed with members of the FKBP family are around 30% in a 100 amino acid overlap (aa 142 to 241) but, intriguingly, this similarity was undetected at the time in the BLAST and FASTA outputs, especially when these were performed with the whole trigger factor sequence, including regions which 'generate' a high background noise, or with the FKBP sequences. However, further sequence analysis revealed that most of the conserved residues are involved in the substrate binding pocket of FKBPs and that residues participating in the maintenance of the hydrophobic core of FKBPs (especially those involved in the regular secondary structures) are found in trigger factor (fig. 15). The high HCA scores (around 80%) deduced from the comparison of trigger factor sequences with members of the FKBP family provide evidence for the conservation of the structural core. Several subtle variations relative to FKBPs which bind FK506 are observed in the catalytic site, such as the canonical Y26 and D37 (FKBP12 numbering, white letters in a grey background in fig. 15) which are conservatively subsituted by F and E respectively in trigger factor. This observation suggested that the functions of the FKBP module of trigger factor may differ slightly from those of canonical FKBP members.

In the meantime, revived interest in trigger factor by different groups has confirmed our PPIase prediction. Indeed, Fischer and coworkers discovered a ribosome-bound prolyl isomerase in *E. coli* and identified it as the trigger factor [46]. Proteolytic fragments encompassing the FKBP-like domain of trigger factor, as well as recombinant forms of this fragment, were shown to

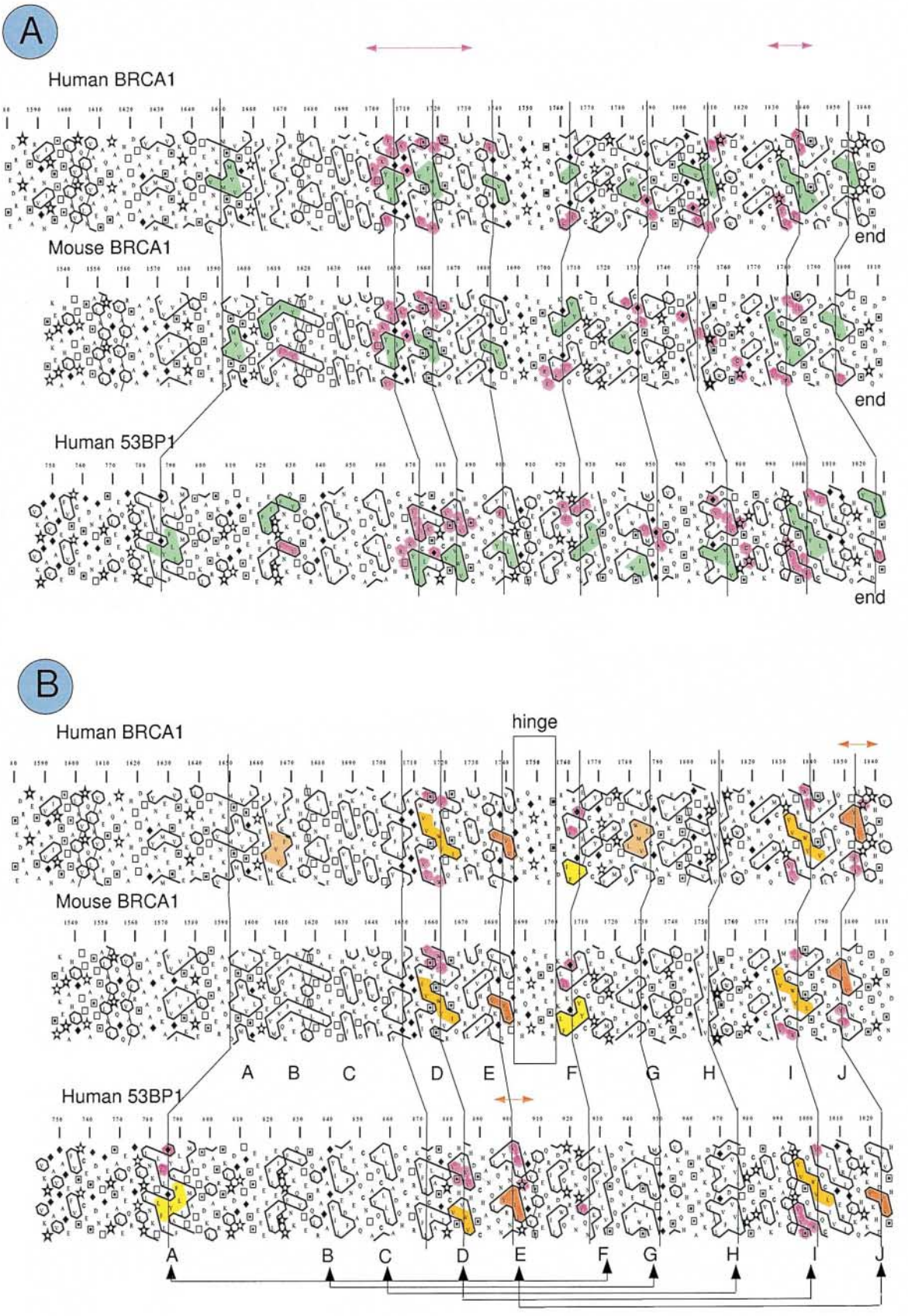
A
Human BRCA1
Mouse BRCA1
Human 53BP1
end
end
end
B
Human BRCA1
hinge
Mouse BRCA1
A B C D E F G H I J
Human 53BP1
A B C D E F G H I J

Figure 13.

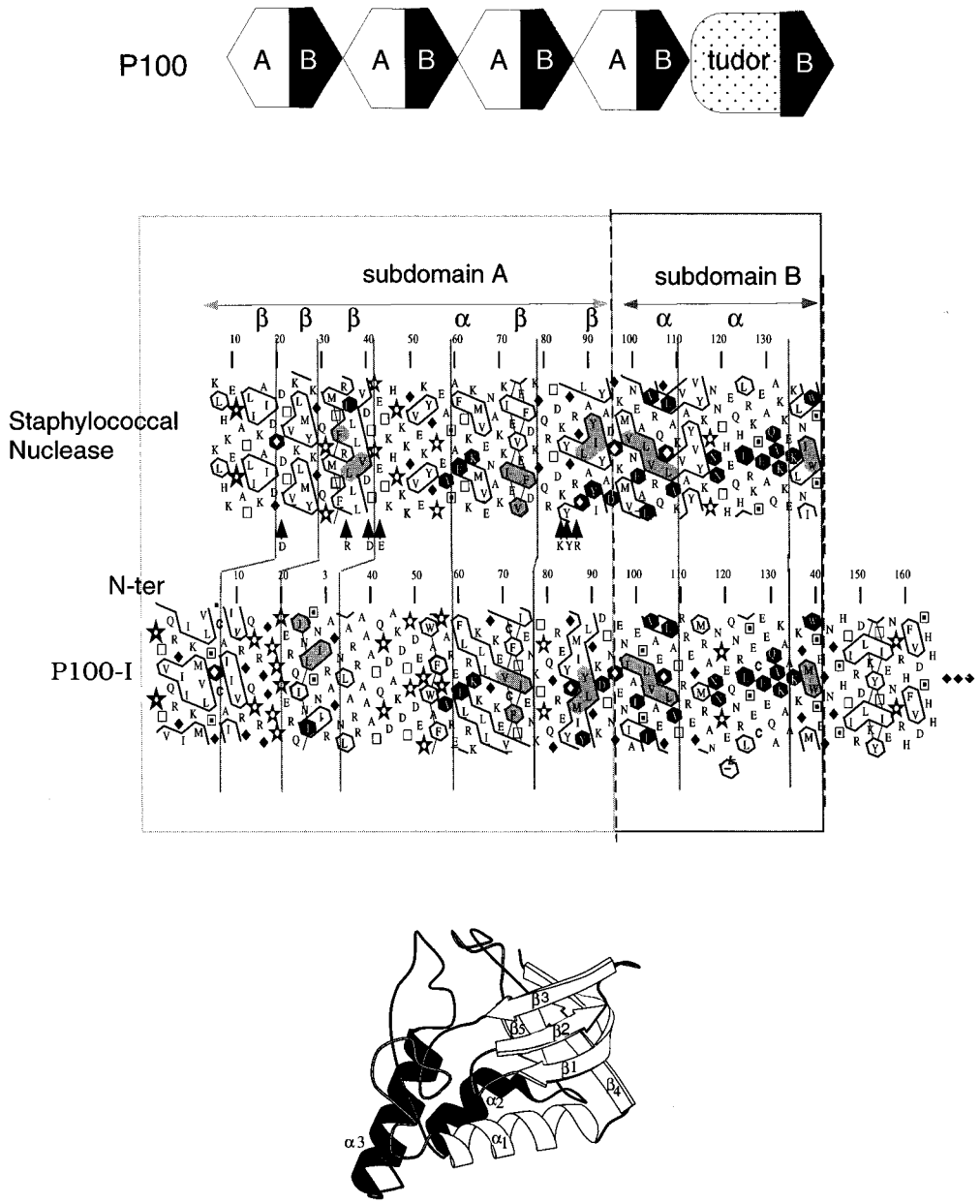


Figure 14. The human EBNA-2 coactivator p100 is composed of a repeat of five domains (top diagram) whose 3D structures can be related to the staphylococcal nuclease fold (bottom). The HCA comparison of the first repeat with the *S. aureus* nuclease is shown (middle), with similar clusters and sequence identities shaded light and dark, respectively. Amino acids which participate in the catalytic activity of SN (arrows) are absent in p100 although the SN fold is conserved. The fifth domain is highly modified, conserving only the second subdomain of the staphylococcal nuclease fold (the two $\alpha$-helices, dark shaded). The first subdomain (not shaded) is replaced by a domain found in multiple copies in the *D. melanogaster* tudor protein.

Figure 13. The C-terminal end of BRCA1 contains a repeated domain shared by a p53-binding protein. (*A*) HCA comparison of the C-terminal domains of BRCA1 (human: Swissprot brc1_human and mouse Swissprot brc1_mouse) and human 53BP1 (Genbank U09477), a p53-binding protein. The two regions of similarity detected by Blast are indicated with pink arrows. The similarity can be extended beyond these two regions, as assessed by cluster similarities (green shaded) and sequence identities (pink shaded). (*B*) Detection of the internal repeat. Similar motifs can be deciphered within the 200 amino acid long domain (same colours), suggesting that motifs A to E (first repeat) could correspond to motifs F to J (second repeat). Moreover, a hinge region is clearly visible between the first and second domains in the BRCA1 sequence. The profile deduced from the conserved motifs has been used to search other members of the family in the sequence databases. The characteristics of the five motifs (A to E) have been described in detail in [26]. The main features of the BRCT domains are found well conserved in the BRCT superfamily, in particular motif B ($\Phi \times (3)\Phi \times (3)$GG, where $\Phi$ is an hydrophobic amino acid), motif C (stretch of three or four hydrophobic amino acids or TH$\Phi\Phi$) and motif D ($\Phi\Phi \times (3)$(W,F)$\Phi \times (2)$(C,S,T)$\Phi$).

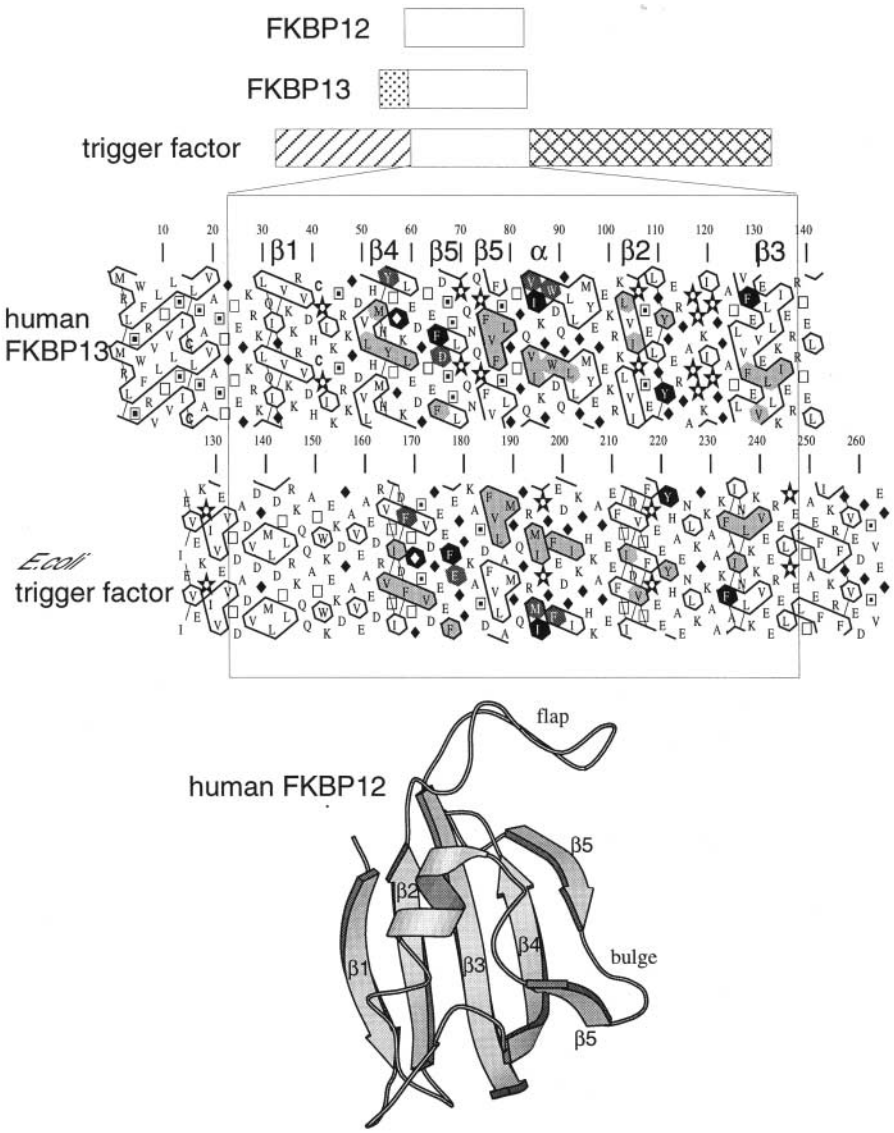


Figure 15. The *E. coli* trigger factor was found to contain a centrally located domain whose 3D structure can be related to the FKBP structure (bottom) found in single or multiple copies in various proteins. Cluster similarities are shown in light shading. Five of the nine amino acids involved in the catalytic site and in FK506 binding are identical (white letters on a black background), the four other are conservatively substituted (white letters on a grey background), suggesting that trigger factor behaviour differs slightly from that of classical FKBP members. Indeed, it was experimentally shown that trigger factor does not bind FK506, although it does have PPIase activity.

retain the full PPIase activity of the intact protein with oligopeptides [47, 48]. On the other hand, trigger factor was shown to interact with nascent proteins (secretory as well as nonsecretory) while they were still associated with the ribosome, indicating that trigger factor would act as a general molecular chaperone [49, 50]. There is also evidence that trigger factor is bound to GroEL in a substrate-dependent manner [51]. Finally, trigger factor appears to have an unusually high folding activity [46], originating from its tight binding to the protein folding substrate and in contrast to its small catalytic constant [52]. The region encircling the FKBP domain may certainly play an important role in the binding ability and chaperone activity of trigger factor.

The breast and ovarian cancer susceptibility antigen BRCA1 is a tumour suppressor gene whose amino acid sequence was reported to share no significant similarity to any other protein, with the exception of its N-terminus which contains a single C3HC4-type zinc finger [53]. When examining its sequence with HCA, a typical globular domain of approximately 200 amino acids was found at the C-terminal end of the protein. This region

appears to be essential, as truncations of the BRCA1 C-terminus result in an inability to suppress breast cancer cell growth [54], and as it acts as a transcriptional transactivator [55]. A BLAST search using this domain as a probe gave an interesting similarity with the C-terminal end of a human p53-binding protein (53BP1) for two short sequence stretches (pink arrows in fig. 13). The similarity can be extended outwards from these overlaps for about 200 amino acids. The potential 3D relationship of these two C-terminal regions was supported by the observation of cluster correspondences between their HCA plots (shaded green in fig. 13), accompanied by several stretches of sequence conservation (shaded pink in fig. 13). Interestingly, it can be seen that the most conserved regions, designated D and I in fig. 13, are strikingly similar, with a pattern $\phi$VxxxWVxx(SC) retrieved in all of them (where $\phi$ is V, I or L and x any residue). Other obvious correlations can be found, for example between E (53BP1) and J (BRCA1), between A (53BP1) and F (mouse BRCA1) or B and G (human BRCA1). These observations led to the hypothesis that these C-terminal domains could consist of a repeat of two similar domains of 100 amino acids. This hypothesis was further strengthened by the observation of a clear hinge region (devoid of any hydrophobic amino acids, shown boxed in fig. 13) between the duplicate domains in BRCA1.

The same conclusion was made by Koonin et al. [56], who, using a signature centred around motifs C and D, highlighted the presence of this domain which they named BRCT (for BRCA1 C-Terminus) in the C-terminal domains of the yeast RAD9 protein and of two hypothetical proteins from human and fission yeast (designated KIAA0170 and SPAC19G10_7, respectively). However, examination of the HCA plots of these proteins reveals four additional BRCT domains in SPAC19G10_7 (thus two copies of the BRCT repeat), not described in [56] because they escape the too-strict signature used to describe the BRCT domain. These domains nonetheless conserve several of the motifs characterizing the BRCT domain.

Iterative searches in the sequence databanks with the different BRCT domains as queries and with HCA as a discriminating tool, allowed us to retrieve a total of 50 copies of this domain in 23 different proteins, including, in addition to BRCA1, 53BP1 and RAD9, XRCC1, RAD4, Ect2, REV1, Crb2, RAP1, terminal deoxynucleotidyl transferases (TdT) and three eukaryotic DNA ligases [26]. The retrieval of the BRCT domain isolated in RAP1 and TdT suggests that it could indeed constitute an autonomous folding unit. The close involvment of most of these proteins in cell cycle regulation and DNA repair emphasizes the potential importance of this domain in the BRCA1 mechanism of action (discussed in [26]). Similar results were published simultaneously by Bork and coworkers [57] who, in contrast to the unified HCA approach, used a wide variety of search methods.

The last three examples illustrate the pitfall that can be encountered when performing exclusively motif and profile searches. Indeed, even highly sensitive position-dependent weight matrices can miss divergent motifs within conserved 2D environments which are recognizable through HCA. The folds of the human p100 protein and the trigger factor belong to families with enzymatic activities and which are characterized, in the motif databases such as PROSITE, by the main features of the more conserved catalytic site. However, both proteins, although conserving the fold of their respective structural families, have lost one of the functions that their canonical members usually perform. Indeed, trigger factor cannot bind FK506 although it still has PPIase activity [46] whereas p100 would cleave DNA, at least through the mechanism adopted by staphylococcal nuclease, although it still binds [37]. Consequently, the assignment of these proteins to well-defined structural families is overlooked owing to the divergence from the functional motif used to describe them. Similarly, most members of the BRCT family were not initially picked by the signature defined by [56] because it was initially constructed on a set of sequences which was too small and therefore not representative of the high divergence encountered in this family.

In these cases, the deduction of the relationships of p100 and trigger factor to the SN and FKBP families, respectively, as well as the description of an extended BRCT family, have come from the accurate analysis of results obtained after iterative scanning of the databases either with single sequences or with profiles deduced from multiple alignments. However, in many cases, 1D methods do not even succeed in picking out the potential candidates, since the 1D motif has evolved much more than the 2D signature, as illustrated in a recent example [58]. This observation stresses the need to develop such tools based on 2D sequence analysis, adding structural information to large-scale database screening.

## HCA false positives

Two predictive studies at very low sequence identity ($<15\%$) and for peculiar structures have led to large discrepancies in the case of nuclear receptors, and more subtle ones for Hsp60-Hsp70 families.

**The hormone (or ligand) binding domains of nuclear receptors.** Ironically, it was the hypothetical connection suggested by HCA between the serpin corticosteroid binding globulin (CBG) sequence and that of the ligand-binding domain of the progesterone receptor (PR) which more than ten years ago encouraged the development of the method. Indeed, as summarized in [59], many structural and biological data have long sup-

ported this hypothesis. The main arguments were the following: (i) CBG and PR both bind the same ligand through globular domains of similar length (~250 amino acids), (ii) both undergo major structural changes associated with their functions, (iii) the sequences of the receptors end at the cleavable serpin reactive loop, (iv) despite a low level of sequence identity (often <10%), the HCA plots of members of the two families show clear similarities supported by the coincidence of two important hydrophobic clusters in the core of the sequences which in serpins are two major $\beta$-strands (A3 and B3, see [59]).

The 3D structures of several ligand-binding domains of nuclear receptors (see e.g. refs 60, 61) have revealed an unexpected and new '$\alpha$-helical sandwich' fold (nearly all $\alpha$) possessing two important buried $\alpha$-helices (H5 and H8) instead of the two suspected $\beta$-strands counterparts of the serpins A3 and B3 segments. Clearly it was this very rare occurrence of two large buried helices, whose HCA shapes are similar to those of typical long $\beta$-strands, which wrongly supported the coherence of the whole domain. Although the serpin and ligand-binding domain folds are unrelated, we observe that, as expected, several local secondary structures have been correctly predicted, i.e. the serpin 'd', 'f', 'g + h' and 'i' helices versus the receptor 'H1', 'H3', 'H9' and 'H10' helices, respectively. Moreover, the receptor mobile C-terminus H11 and H12 helices which exhibit a 'mousetrap' mechanism occupy sequence positions near that of the serpin mobile active loop in the proposed alignment. While HCA proposed the $\alpha + \beta$ serpin model, several other false $\alpha + \beta$ models have also been published [62–65]. An interesting recent study shows how the multivariate analysis of amino acid composition could help HCA in such borderline studies, efficiently separating the two families compared here [66].

**Hsp60 and Hsp70 chaperone proteins.** The general chaperone proteins Hsp60 (GroEl) and Hsp70 share several common features: (i) their polypeptide chains are of similar length and protect and/or process the polypeptide chain of other proteins during folding/unfolding steps, and (ii) both use cycles of ATP/ADP transformation to achieve their functions. However, Hsp70 works as a monomer or a dimer while Hsp60 constitutes a large tunnelled heptameric structure associated with the heptameric but smaller ring of the co-chaperonin GroEs. The 3D structure of Hsc70 (a member of the Hsp70 family) was determined first [67], while that of GroEl is more recent [68].

Hsp70 shares 3D similarities with actin and hexokinase within its ATP-binding domain which can be divided into four subdomains, the ATP molecule being located in a cleft between them. These unexpected structural similarities have long escaped sequence analysis due to the very weak level of sequence identity (less than 10%) [69].

While waiting for the resolution of the GroEl 3D structure, our attention was caught by several HCA similarities between the sequences of Hsp70 ATP-binding domains and those of the Hsp60 family, suggesting a possible 3D similarity between the two [70]. When the structure of GroEl was solved it was, at a first glance, quite disappointing. Clearly, the Hsp70 and Hsp60 structures are not directly superimposable. However, the path of the polypeptide chains between the two large lobes encircling ATP in these proteins are curiously similar although the secondary structures are essentially different. Consequently, the overall shape of GroEl shares similarities with several members of the actin/hexokinase family including Hsc70 and poses an intriguing question about a possible convergent or divergent evolution between ATP-dependent molecular motors, probably also including Hsp90. In this respect, the above study can be considered as a partial HCA false positive. In the context of this study, we speculated further about chaperone functions and their mode of action [71].

**Statistical significance of alignments in the twilight zone.** When two or more sequences are put into correspondence through HCA at a low level of sequence identity (15–10%), the corresponding amino acid alignment should be carefully established by human processing and editing starting from the anchor points recognized on the plots. This procedure rejects insertions and deletions in the loop regions situated mainly between clusters. Often loops cannot be aligned, as their length and composition differ totally. Based on this 1D alignment, identity scores, similarity scores (calculated with an appropriate substitution matrix [72]) and HCA scores can then be calculated. The HCA score (percentage) is the ratio between the number of positions both occupied by hydrophobic amino acids in the two compared sequences and the total number of hydrophobic positions. HCA values above 60% are often observed for structurally related domains [1].

The evaluation of the statistical significance of alignments is always tricky at low levels of sequence identity. The central question is to estimate if the observed alignment may be expected to occur purely by chanceor, in other words, if it significantly differs from background noise.

To this end, identity, similarity and HCA scores deduced from the pairwise comparisons can be compared through Z-scores to the distribution of scores obtained after alignment of the first sequence ($i$) versus shuffled versions of the second one ($j$) (typically 1000 to 10,000 times). This method, first introduced by Doolittle [73] and discussed by Lipman and coworkers [74] in the case of nucleic acids, has been widely used to assess the significance of global alignments [75].

The Z-score $Z_{ij}$ obtained from the comparison of the sequence i to the sequence j corresponds to the differ-

ence between the observed score ($Q_{ij}$) and the mean score of the random distribution ($Q_m$), expressed relative to the standard deviation of the distribution ($\sigma$).

$$Z_{ij} = (Q_{ij} - Q_m)/\sigma$$

The random scores do not have a pure normal distribution and, in particular, they often present a tail at large values of the scores. A Z-score value above 6 standard deviations is generally taken as a reliable proof of relatedness. An exhaustive study performed on the yeast genome and correlating the calculated random distribution for local alignments to the extreme value distribution used to describe the optimal scores of ungapped subalignments, as expressed by the Gumble law [76], has confirmed the above threshold value between 6 and 8 for the significance level (J.-P. Comet, J.-C. Aude, E. Clément, A. Hénaut, J.-L. Risler, P. Slonimski et al., unpublished observations). Actually the length of the sequence seems to affect the significance threshold value directly. Several observations indicate that short related sequences present smaller Z-scores (principally due to larger standard deviations) than longer sequences for similar score levels. In other words, the rule of thumb is: the shorter the sequence, the smaller the Z-score. Thus, it should be kept in mind that the consideration of Z-scores in assessing the authenticity of a relationship is also a function of the sequence length. This has not been explained on a theoretical basis (as far as we know), but it could be due to the size and composition of the amino acid alphabet as the number of possible combinations for short sequences is smaller than for longer ones.

The three Z-scores have been combined in a term called Z3, which is an indication of sequence relatedness versus random distributions.

Taking into account the information included in a family of sequences instead of that encoded in a single sequence significantly increases the discriminating power of Z-scores in distinguishing genuine relationships. In this respect, scores and random scores are calculated between a 'family' ($I$) and the sequence ($j$) which is thought to belong to the family. The family ($I$) sequence can be represented by an $n \times 21$ matrix where n is the total length of the compared sequences (including gaps) and 21 corresponds to the 20 amino acids + one position reserved for gaps. This 'profile-like' procedure is more sensitive than that based on the consideration of a consensus sequence. In particular, it makes it possible to distinguish false positives having similar levels of sequence identity from authentic members of a family, especially on the basis of the HCA Z-score.

**HCA software note.** Classical HCA plots (black/white or colour) can easily be drawn through our Web server at the URL http://www.lmcp.jussieu/~mornon. This free service is called drawhca. Since the HCA plots are PostScript files, they can be imported within a graphical software and then processed. We currently use IslandDraw (Island Software, The Netherlands) for Unix workstations and CorelDraw (Corel Corporation, Canada) for Windows 95.

Useful additional software is also accessible on our Web server:

- MulBLAST [31]:
  http://www.lmcp.jussieu.fr/~mornon in the section MulBLAST
  http://www.lmcp.jussieu.fr/~labesse
- Visual BLAST and Visual FASTA [30]:
  http://www.lmcp.jussieu.fr/~mornon in the section Visual BLAST
  http://www.lmcp.jussieu.fr/~durand in the section Softwares
- Tzscore: a program devoted to the assessment of alignments in and below the twilight zone will soon be available at the URL http://www.lmcp.jussieu.fr/~mornon in the section TZscore.
- PSEA [15]: an efficient new program to assign $\alpha$, $\beta$ or coil conformations of each amino acid from the protein C$\alpha$ 3D coordinates:
  http://www.lmcp.jussieu.fr/~mornon in the section PSEA
  http://www.lmcp.jussieu.fr/~labesse

## Perspectives

The bidimensional HCA plot of a sequence or a family of sequences allows an overall appreciation of many structural features of proteins at a glance. Subsequently, a methodical approach often helps to overcome several pitfalls of 1D sequence analysis and sequence comparison at low levels of sequence identity. The overall efficiency of the method, however, depends on multiple factors, several of which are human ones. Indeed, HCA has proven to be well suited to the efficiency of human information processing and obviously requires a balance between rigor and intuition, as in many scientific fields. Therefore, full automation of sequence comparison at very low levels of sequence identity appears still out of reach for the moment. However, a limited automation procedure allowing a complete 'prescanning'of large sequence databases through HCA seems feasible, considering the large amount of data accumulated in recent years. This aim at least partially combines interesting perspectives concerning a better understanding of protein folding through an extensive analysis of many structural bricks centred on hydrophobic clusters. Joining together amino acids which are separated by intermediate distances on the sequence makes it possible to skip over a step in the not yet deciphered path leading from sequence to structure and, subsequently, to protein function.

In this context, the peculiar properties of hydrophobic clustering on helical transpositions and their associated criteria of hydrophobic compactness open new perspec-

tives using hyperhelices in the 3D plot/ 4D space which are now under study in our laboratory. In this respect, HCA – **H**ydrophobic **C**luster **A**nalysis – is also **H**elical **C**lustering **A**nalysis.

*Acknowledgements*. The authors are pleased to acknowledge the enthusiastic support of many colleagues around the world when they have explored their sequence data through HCA. Special thanks go to Guy Bourat for many stimulating and thoughtful discussions concerning the HCA world and to all our coworkers in the field. The authors also acknowledge the continuous support of Paul Hossenlopp concerning HCA training through the *CNRS Formation permanente*. HCA development over the ten past years has been supported by public and industrial research, including CNRS, INSERM, Universités P6 and P7, Rhône-Poulenc Rorer and Roussel-Uclaf and several research programs such as Sidaction, CM2AO, Vaincre les Maladies Lysosomales, and the EEC Biotech program.

1 Lemesle-Varloot L., Henrissat B., Gaboriaud C., Bissery V., Morgat A. and Mornon J.-P. (1990) Hydrophobic cluster analysis: procedures to derive structural and functional information from 2D representation of protein sequences. Biochimie **72:** 555–574
2 Chothia C. (1992) One thousand families for the molecular biologist. Nature **357:** 543–544
3 Li H., Helling R., Tang C. and Wingreen N. (1996) Emergence of preferred structures in a simple model of protein folding. Science **273:** 666–669
4 Kardar M. (1996) Which came first, protein sequence or structure? Science **273:** 610–611
5 Altschul S., Gish W., Miller W., Myers E. and Lipman D. (1990) Basic local alignment search tool. J. Molec. Biol. **215:** 403–410
6 Pearson W. R. and Lipman D. J. (1988) Improved tools for biological comparison. Proc. Natl Acad. Sci. USA **85:** 2444–2448
7 Koonin E. V., Tatusov R. L. and Rudd K. E. (1996) Protein sequence comparison at genome scale. Meth. Enzymol. **266:** 295–322
8 Bork P. and Gibson T. J. (1996) Searching motif and profile searches. Meth. Enzymol. **266:** 162–184
9 Gaboriaud C., Bissery V., Benchetrit T. and Mornon J.-P. (1987) Hydrophobic cluster analysis. An efficient new way to compare and analyse amino-acid sequences. FEBS Lett. **224:** 149–155
10 Woodcock S., Mornon J.-P. and Henrissat B. (1992) Detection of secondary structure elements in proteins by Hydrophobic Cluster Analysis. Prot. Eng. **5:** 629–635
11 Dunhill P. (1968) The use of helical net diagrams to represent protein sequences. Biophys. J. **8:** 865–875
12 Lim V. I. (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J. Molec. Biol. **88:** 857–872
13 Lim V. A. (1974) Algorithm for prediction of $\alpha$-helical and $\beta$-structural regions in globular proteins. J. Molec. Biol. **88:** 873–894
14 Hobohm U. and Sander C. (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci. **1:** 409–417
15 Labesse G., Colloc'h N., Pothier J. and Mornon J.-P. (1997) P-SEA: a program for secondary structure assignment from C$\alpha$. Comp. Appl. Biosci. **13:** 291–295
16 Bourat G., Thoreau E. and Mornon J.-P. (1994) 2D-helical hydrophobic clusters: statistics and morphology. J. Pharmacol. Belg. **49:** 226–235
17 Vazquez S., Thomas C., Lew R. A. and Humphreys R. E. (1993) Favored and suppressed patterns of hydrophobic and nonhydrophobic amino acids in protein sequences. Proc. Natl Acad. Sci. USA **90:** 9100–9104
18 West M. W. and Hecht M. H. (1995) Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. Protein Sci. **4:** 2032–2039
19 Pauling L., Corey R. B. and Branson H. R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl Acad. Sci. USA **37:** 205–211
20 Pan K.-M., Baldwin M. A., Nguyen J., Gasset M., Serban A., Groth D. et al. (1993) Conversion of $\alpha$-helices into $\beta$-sheets features in the formation of the scrapie prion proteins. Proc. Natl Acad. Sci. USA **90:** 10962–10966
21 Callebaut I., Tasso A., Brasseur R., Burny A., Portetelle D. and Mornon J.-P. (1994) Common prevalence of alanine and glycine in mobile reactive centre loops of serpins and in viral fusion peptides. Do prions possess a fusogenic peptide? J. Comp. Aided Molec. Design **8:** 175–191
22 Forloni G., Angeretti N., Chiesa R., Monzani E., Salmona M., Bugiani O. et al. (1993) Neurotoxicity of a prion protein fragment. Nature **362:** 543–546
23 De Gioia L., Selvaggini C., Ghibaudi E., Diomede L., Bugiani O., Forloni G. et al. (1994) Conformational polymorphism of the amyloidogenic and neurotoxic peptide homologous to residues 106–126 of the prion protein. J. Biol. Chem. **269:** 7859–7862
24 Heller J., Kolbert A. C., Larsen R., Ernst M., Bekker T., Baldwin M. et al. (1996) Solid-state NMR studies of the prion protein H1 fragment. Protein Sci. **5:** 1655–1661
25 Riek R., Hornemann S., Wider G., Billeter M., Glockshuber R. and Wüthrich K. (1996) NMR structure of the mouse prion protein domain PrP(121–231). Nature **382:** 180–182
26 Callebaut I. and Mornon J.-P. (1997) From BRCA1 to RAP1: a widespread BRCT module closely associated to DNA repair. FEBS Lett. **400:** 25–30
27 Thoreau E., Petridou B., Kelly P. A., Djiane J. and Mornon J.-P. (1991) Structural symmetry of the extra-cellular domain of the Cytokine/Growth hormone/Prolactin receptor family and interferon receptors revealed by Hydrophobic Cluster Analysis. FEBS Lett. **282:** 26–31
28 de Vos A. M., Ultsch M. and Kossiakoff A. A. (1992) Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. Science **255:** 306–312
29 Altschul S. F., Bokusi M. S., Gish W. and Wotton J. C. (1994) Issues in searching molecular sequence databases. Nature Genet. **6:** 119–129
30 Durand P., Canard L. and Mornon J.-P. (1997) Visual Blast and Visual FastA: graphic workbenches for interactive analysis of full Blast and FastA outputs under Microsoft Windows 95/NT. Comput. Appl. Biosci. **13:** 407–413
31 Labesse G. (1997) MulBlast 1.0: a multiple alignment of BLAST output to boost protein sequence similarity analysis. Comput. Appl. Biosci. **12:** 463–467
32 Tatusov R. L., Altschul S. F. and Koonin E. V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc. Natl Acad. Sci. USA **91:** 12091–12095
33 Hegyi H. and Pongor S. (1993) Predicting potential domain homologies from FASTA search results. Comput. Appl. Biosci. **9:** 371–372
34 Tatusov R. L. and Koonin E. V. (1994) A simple tool to search for sequence motifs that are conserved in BLAST outputs. Comput. Appl. Biosci. **10:** 457–459
35 Worley K. C., Wiese B. A. and Smith R. F. (1995) BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information ressources into sequence similarity search results. Genome Res. **5:** 173–184
36 Turkenburg J. P. and Dodson E. J. (1996) Modern developments in molecular replacement. Curr. Opin. Struct. Biol. **6:** 604–610
37 Tong X., Drapkin R., Yalamanchill R., Mosialos G. and Kieff E. (1995) The Epstein-Barr virus nuclear protein 2 acidic domain forms a complex with a novel cellular coactivator that can interacr with TFIIE. Molec. Cell. Biol. **15:** 4735–4744
38 Callebaut I. and Mornon J.-P. (1997) The human EBNA-2 coactivator p100: multidomain organization and relationship

to the staphylococcal nuclease fold and to the tudor protein involved in *Drosophila melanogaster* development. Biochem. J. **321:** 125–132
39 Murzin A. G. (1993) OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. **12:** 861–867
40 Ponting C. P. (1997) Tudor domains in proteins that interact with RNA. Trends Biochem. Sci. **22:** 51–52
41 Crooke E. and Wickner W. (1987) Trigger factor: a soluble protein that folds proOmpA into a membrane-assembly-competent form. Proc. Natl Acad. Sci. USA **84:** 5216–5220
42 Crooke E., Brundage L., Rice M. and Wickner W. (1988) ProOmpA spontaneously folds in a membrane assembly competent state which trigger factor stabilizes. EMBO J. **7:** 1831–1835
43 Crooke E., Guthrie B., Lecker S., Lill R. and Wickner W. (1988) ProOmpA is stabilized for membrane translocation by either purified E. coli trigger factor or canine signal recognition particle. Cell **54:** 1003–1011
44 Lecker S., Lill R., Ziegelhoffer T., Georgopoulos C., Bassford P. J., Kumamoto C. A. et al. (1989) Three pure chaperone proteins of *Escherichia coli* – SecB, trigger factor and GroEL – form soluble complexes with precursor proteins in vitro. EMBO J. **8:** 2703–2709
45 Callebaut I. and Mornon J.-P. (1995) Trigger factor, one of the *Escherichia coli* chaperone proteins, is an original member of the FKBP family. FEBS Lett. **374:** 211–215
46 Stoller G., Rücknagel K. P., Nierhaus K. H., Schmid F. X., Fischer G. and Rahfeld J.-U. (1995) A ribosome-associated peptidyl-prolyl cis-trans isomerase identified as the trigger factor. EMBO J. **14:** 4939–4948
47 Stoller G., Tradler T., Rücknagel J.-U. and Fischer G. (1996) A 11.8 kDa proteolytic fragment of the *E. coli* trigger factor represents the domain carrying the peptidyl-prolyl cis-trans isomerase activity. FEBS Lett. **384:** 117–122
48 Hesterkamp T. and Bukau B. (1996) Identification of the prolyl isomerase domain of *Escherichia* trigger factor. FEBS Lett. **385:** 67–71
49 Valent Q. A., Kendall D. A., High S., Kusters R., Oudega B. and Luirink J. (1995) Early events in proprotein recognition in *E. coli*: interaction of SRP and trigger factor with nascent polypeptides. EMBO J. **14:** 5494–5505
50 Hesterkamp T., Hauser S., Lutcke H. and Bukau B. (1996) *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. Proc. Natl Acad. Sci. USA **93:** 4437–4441
51 Krandror O., Sherman M., Rhode M. and Goldberg A. L. (1995) Trigger factor is involved in GroEL-dependent protein degradation in *Escherichia coli* and promotes binding of GroEL to unfolded proteins. EMBO J. **14:** 6021–6027
52 Scholtz C., Stoller C., Zarnt T., Fischer G. and Schmid F. X. (1997) Cooperation of enzymatic and chaperone functions of trigger factor in the catalysis of protein folding. EMBO J. **16:** 54–58
53 Miki Y., Swensen J., Shattuck-Eidens D., Futreal P. A., Harshman K., Tavtigian S. et al. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science **66:** 66–71
54 Holt J. T., Thompson M. E., Szabo C., Robinson-Benion C., Arteaga C. L., King M.-C. et al. (1996) Growth retardation and tumor inhibition by BRCA1. Nature Genet. **12:** 298–302
55 Chapman M. S. and Verma I. M. (1996) Transcriptional activation by BRCA1. Nature **382:** 678–679
56 Koonin E. V., Altschul S. F. and Bork P. (1996) BRCA1 protein products: functional motifs. Nature Genet. **13:** 266–267
57 Bork P., Hofmann K., Bucher P., Neuwald A. F., Altschul S. F. and Koonin E. V. (1997) A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. FASEB J. **11:** 68–76
58 Callebaut I., Courvalin J.-C., Worman H. J. and Mornon J.-P. (1997) Hydrophobic Cluster Analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. Biochem. Biophys. Res. Commun. **235:** 103–107
59 Mornon J.-P., Thoreau E., Rowlands D., Callebaut I. and Moreau G. (1994) Putative dimeric organization of nuclear receptor hormone-binding domains, as suggested by Hydrophobic Cluster Analysis. C. R. Acad. Sci. Paris, Life Sci. **317:** 597–606
60 Renaud J.-P., Rochel N., Ruff M., Vivat V., Chambon P., Gronemeyer H. et al. (1995) Crystal structure of the RAR-$\gamma$ ligand-binding domain bound to all-*trans* retinoic acid. Nature **378:** 681–689
61 Wagner R. L., Apriletti J. W., McGrath M. E., West B. L., Baxter J. D. and Fletterick R. J. (1995) A structural role for hormone in the thyroid hormone receptor. Nature **378:** 690–697
62 Lewis D. F. V. and Lae B. G. (1993) Interaction of some peroxisome proliferators with the mouse liver peroxisome proliferator-activated receptor (PPAR) a molecular modelling and quantitative structure-activity relationship (QSAR) study. Xenobiotica **23:** 73–96
63 Goldstein R. A., Katzenellenbogen J. A., Luthey-Schulten Z. A., Seielstad D. A. and Wolynes P. G. (1993) Three-dimensional model for the hormone binding domains of steroid receptors. Proc. Natl Acad. Sci. USA **90:** 9949–9953
64 Höltje H. D. and Dall N. (1993) A molecular modelling study on the hormone binding site of the estrogen receptor. Pharmazie **48:** 243–249
65 McPhie P., Parkison C., Lee B. K. and Cheng S. Y. (1993) Structure of the hormone binding domain of human $\beta$1 thyroid hormone nuclear receptor. Is it an $\alpha/\beta$ barrel? Biochemistry **32:** 7460–7465
66 Ojasoo T. and Doré J.-C. (1996) Taxonomy of nuclear receptors and serpins by multivariate analysis of amino-acid composition. J. Steroid Biochem. Molec. Biol. **58:** 167–181
67 Flaherty K. M., DeLuca-Flaherty C. and McKay D. (1990) Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein. Nature **346:** 623–628
68 Braig K., Otwinowski Z., Hedge R., Boisvert D. C., Joachimiak A., Horwich A. L. et al. (1994) The crystal structure of the bacterial chaperonin GroEl at 2.8 Å. Nature **371:** 578–586
69 Flaherty K. M., McKay D., Kabsch W. and Holmes K. (1991) Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. Proc. Natl Acad. Sci. USA **88:** 5041–5045
70 Callebaut I., Catelli M. G., Portetelle D., Burny A., Baulieu E. E. and Mornon J.-P. (1994) Structural similarities between chaperone molecules of the HSP60 and HSP70 families deduced from Hydrophobic Cluster Analysis. FEBS Lett. **342:** 242–248
71 Callebaut I., Catelli M. G., Portetelle D., Meng X., Cadepond F., Burny A. et al. (1994) Redox mechanism for the chaperone activity of the heat shock protein HSP60, 70 and 90 as suggested by Hydrophobic Cluster Analysis: hypothesis. C.R. Acad. Sci. Paris, Life Sci. **317:** 721–729
72 Henikoff S. (1996) Scores for sequence searches and alignments. Curr. Opin. Struct. Biol. **6:** 353–360
73 Doolittle R. F. (1981) Similar amino acid sequences: chance or common ancestry. Science **214:** 149–159
74 Lipman D. J., Wilbur W. J., Smith T. F. and Waterman M. S. (1984) On the statistical significance of nucleic acid similarities. Nucl. Acids Res. **12:** 215–226
75 Landès C., Hénaut A. and Risler J.-L. (1992) A comparison of several similarity indices used in the classification of protein sequences. A multivariate analysis. Nucl. Acids Res. **20:** 3631–3637
76 Karlin S. and Altschul S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl Acad. Sci. USA **87:** 2264–2268