PROTEINS: Structure, Function, and Bioinformatics 67:695–708 (2007)

Periodic Distributions of Hydrophobic Amino Acids Allows the Definition of Fundamental Building Blocks to Align Distantly Related Proteins

J. Baussand,[1] C. Deremble,[2] and A. Carbone[1]*

[1]*Génomique Analytique, INSERM UMRS511, Université Pierre et Marie Curie-Paris 6, 91, Bd de l'Hôpital, 75013 Paris, France*

[2]*Laboratoire de Biochimie Théorique, IBPC, 13, rue Pierre et Marie Curie, 75005 Paris, France*

ABSTRACT **Several studies on large and small families of proteins proved in a general manner that hydrophobic amino acids are globally conserved even if they are subjected to high rate substitution. Statistical analysis of amino acids evolution within blocks of hydrophobic amino acids detected in sequences suggests their usage as a basic structural pattern to align pairs of proteins of less than 25% sequence identity, with no need of knowing their 3D structure. The authors present a new global alignment method and an automatic tool for Proteins with HYdrophobic Blocks ALignment (PHYBAL) based on the combinatorics of overlapping hydrophobic blocks. Two substitution matrices modeling a different selective pressure inside and outside hydrophobic blocks are constructed, the Inside Hydrophobic Blocks Matrix and the Outside Hydrophobic Blocks Matrix, and a 4D space of gap values is explored. PHYBAL performance is evaluated against Needleman and Wunsch algorithm run with Blosum 30, Blosum 45, Blosum 62, Gonnet, HSDM, PAM250, Johnson and Remote Homo matrices. PHYBAL behavior is analyzed on eight randomly selected pairs of proteins of <30% sequence identity that cover a large spectrum of structural properties. It is also validated on two large datasets, the 127 pairs of the Domingues dataset with <30% sequence identity, and 181 pairs issued from BAliBASE 2.0 and ranked by percentage of identity from 7 to 25%. Results confirm the importance of considering substitution matrices modeling hydrophobic contexts and a 4D space of gap values in aligning distantly related proteins. Two new notions of local and global stability are defined to assess the robustness of an alignment algorithm and the accuracy of PHYBAL. A new notion, the SAD-coefficient, to assess the difficulty of structural alignment is also introduced. PHYBAL has been compared with Hydrophobic Cluster Analysis and HMMSUM methods. Proteins 2007;67:695–708.** © 2007 Wiley-Liss, Inc.

Key words: sequence alignment; evolution; remote proteins; protein homology; substitution matrices; gaps; secondary structures; hydrophobic blocks

INTRODUCTION

Proteins sharing more than 30% of sequence identity have a high probability to also share the same fold.[1] Thus, as fold and function of a protein generally have an intimate relationship, strong sequence homology is exploited by conventional sequence comparison methods to detect these similarities, reconstruct families of functionally related proteins, and accomplish annotation of genomes. Unfortunately, the complete sequencing of several organisms differing in physiology, habitat, and genetics brought to light how weak this approach to annotation can be: for some genomes, as the malaria parasite *Plasmodium falciparum*, more than half of the genes remain functionally unknown. Alignment of pairs of sequences with <30% identity is known to be difficult[2]: there are several examples of proteins where widely used methods like BLAST[3,4] and CLUSTALW[5] together with suitable choice of score matrices and gap values do not detect any homology, but possibly issue lists of pairs of proteins with low scores or high e-values that once screened with finer approaches might reveal 10–15% of sequence identity, same structure and same function.[6,7]

Conventional sequence comparison methods use general empirical models of proteins evolution represented by substitution matrices[8–10] and gap penalties. However,

Abbreviations: aa, amino acid; bGEP, Gap Extension Penalties within blocks; bGOP, Gap Opening Penalties within blocks; br, buried residues; CAP, number of correctly aligned pairs; GEP, Gap Extension Penalties outside blocks; GOP, Gap Opening Penalties outside blocks; hb, hydrophobic block; HCA, Hydrophobic Cluster Analysis; IHBM, Inside Hydrophobic Blocks Matrix; OHBM, Outside Hydrophobic Blocks Matrix; PCA, Principal Component Analysis; PCAP, percentage of correctly aligned pairs; rss, regular secondary structure; SAD, Structural Alignment Difficulty; thb, true hydrophobic block.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Alessandra Carbone, Génomique Analytique, INSERM UMRS511. Université Pierre et Marie Curie-Paris 6, 91, Bd de l'Hôpital, 75013 Paris, France. E-mail: Alessandra.Carbone@lip6.fr

Received 23 June 2006; Revised 28 September 2006; Accepted 2 November 2006

Published online 13 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21319

these models omit two fundamental aspects of evolution that explain their inefficiency in aligning proteins of low homology. First, residue positions are not equally important in a protein sequence and they are not subject to the same rate of mutation.[11] Second, because of specific evolutionary pressures, the type of substitution is strongly related to the functional and structural role of the residue: evidence of differential evolution has been shown between residues belonging to either regular secondary structures (rss) or coil, and also between buried residues (br) and exposed ones.[11–13] Thus, it is essential that substitution matrices and gap penalties take into account residues positions and in particular their structural environment.[14,15] Several substitution matrices have been constructed taking into account residues location in rss and exposure to solvent, and have been successfully applied to folding recognition.[13,16,17] Also, new pairwise alignment methods as HMMSUM[18] are based on a prediction of conserved structural regions and adapt their substitution matrices accordingly along the alignment. These models support the idea that structural information helps aligning distantly related proteins.

Clusters of hydrophobic amino acids that are in contact once the protein is folded, appear crucial to the folding process and to the formation of rss.[19–26] On the basis of this observation, groups of hydrophobic amino acids definable on a sequence and usually corresponding to facets of rss lying in the protein internal core[27] enabled to show weak homology (15–25%) of several families of proteins by manually aligning sequences and their hydrophobic structures using the Hydrophobic Cluster Analysis (HCA) method.[28,29] The methodology of protein sequence analysis and alignment of distantly related proteins that we present here uses blocks of hydrophobic residues (hb) defined as one-dimensional (1D) variants of the notion of hydrophobic cluster, by detecting closely located hydrophobic residues in a sequence and including intercalating nonhydrophobic ones. It turns out that hb are highly correlated to rss and to br and consequently, they can be considered as structural patterns extracted from sequences, with no need of knowing the 3D structure of the proteins to be aligned.

On the basis of hb, we define a new automatic method, called Proteins with HYdrophobic Blocks ALignment [PHYBAL (http://www.ihes.fr/∼carbone/data.htm)], for pairwise sequence alignment. Our tool identifies hb in one dimension, and integrates this structural information in an alignment procedure by imposing to the Needleman and Wunsch[30] global (and global-local) alignment algorithm further structural conditions suggested by hb. Two substitution matrices modeling a different selective pressure within and without hb are constructed, the Inside Hydrophobic Blocks Matrix (IHBM) and the Outside Hydrophobic Blocks Matrix (OHBM), and a 4D space of gap values is explored.

We analyze in detail eight randomly selected protein pairs with at most 26% sequence identity and different structural characteristics, and validate our approach by testing alignment performance on two large datasets of protein pairs with less than 30% of sequence identity, extracted from BAliBASE 2.0 and Domingues datasets. We show that PHYBAL improves response on both datasets compared with its alignment algorithm run with commonly used substitution matrices. Results confirm the importance of considering substitution matrices modeling evolutionary pressure within and without hb and of considering a 4D space of gap values in aligning distantly related proteins. In particular, PHYBAL working with IHBM and OHBM displays a stable behavior of the alignment procedure on close variations of gap values. The stability property is desirable since high divergence implies exact values of gap penalties to be important for success.[5] Two new notions of local and global stability are defined to assess the robustness of an alignment algorithm and the accuracy of PHYBAL. A new notion, the Structural Alignment Difficulty (SAD)-coefficient, to assess the difficulty of structural alignment is also introduced.

PHYBAL approach has been compared with two alignment methods based on the idea that considering structural contexts in sequences should help for a better alignment, the manual HCA method and the automatic HMMSUM method.

MATERIAL AND METHODS

Definitions of Hydrophobic Blocks

A significative periodic distribution of hydrophobic amino acids along a protein sequence can be observed in α-helices and β-sheets. Periods seem to be dependent on the rss location in the folded protein[7]: α-helices lying on the protein surface display contacts among hydrophobic residues at distance 3 or 4; β-sheets lying on the protein surface display contacts among hydrophobic residues at distance 2; and secondary structures lying within an hydrophobic core display chains of consecutive hydrophobic residues, that is at distance 1. On the basis of these simple combinatorial patterns, one can define formal rules to detect *hydrophobic blocks* on a sequence, without the knowledge of the 3D structure, by reading the periodicity of hydrophobic residues supposed to be adjacent in three dimensions.

Hydrophobic blocks

Formally, we define the following combinatorial rule: reading a sequence from left to right, for each amino-acid k_0, two situations can occur: (a) if k_0 is not hydrophobic then continue reading; (b) if k_0 is hydrophobic then consider its four consecutive neighbors on the right. If there is a minimal index $1 \leq j \leq 4$ such that k_j is a proline, then consider the maximum index $0 \leq l < j$ such that k_l is hydrophobic. Otherwise consider the maximum index $0 \leq l \leq 4$ such that k_l is hydrophobic. If $l = 0$ then continue reading the sequence. Otherwise consider the following two cases: if k_0 is not yet in a block, define a block as being constituted by all aa from k_0 to k_l included. If k_0 belongs to a block C, then extend C by

adding to it all aa from k_0 to k_l included. Continue the reading of the sequence by setting $k_0 = k_l$.

By definition, note that certain hb might be broken by the presence of a proline sitting between two neighboring hydrophobic residues.

Hydrophobic amino acids

We consider as hydrophobic the following aa: valine (V), isoleucine (I), leucine (L), phenylalanine (F), methionine (M), tryptophane (W), and tyrosine (Y).[28] Prolines (P) are considered as "hydrophobic blocks breakers" since they tend to initiate sudden variations in the direction of the aa chain in tree dimensions.

Definition of Three Datasets of Structural Alignments Used to Construct Substitution Matrices

We considered the database HOMSTRAD[31] of protein families and structural alignments from which we removed those proteins with undefined rss in their PDB files. Also, we removed or slightly modified (at most three residues) sequences from the HOMSTRAD protein families that differed from the corresponding PDB files. For each protein family issued from HOMSTRAD, we retained a subset of proteins such that (1) protein pairs have less than 30% identity, (2) each family contains at least three sequences, and (3) are classified to have the same common ancestor (according to the HOMSTRAD classification discussed in www-cryst.bioc.cam.ac.uk/~homstrad/Doc/Info.html). Strictly speaking, we count sequences instead of proteins since in certain cases, HOMSTRAD aligns paralogous domains lying in the same protein. It gives us 144 families with 613 sequences coming from 523 proteins, with 30 α/β, 37 α + β, 21 all α, 27 all β, 8 αβ-barrel, 11 multidomain, 5 small, 4 small disulphide, 1 membrane bound all β. To reach a wide variability in residue substitution, we favored a large number of small families and considered 86 with 3 sequences; 30 with 4, 19 with $\geq$5 and $\leq$10 sequences, and only 9 with >10 but $\leq$14 sequences. Protein pairs are 1426. See Supplementary Table I.

Because of the different behavior of structural alignment approaches, three datasets of structural alignments have been constructed from this selected set of protein families using the alignments proposed in HOMSTRAD and the server ProSup. One dataset is issued from HOMSTRAD, which collects alignments of protein families realized with FUGUE[32] together with MNYFIT, COMPARER, and STAMP.[33] FUGUE is a tool performing structural alignment based on sequence information and on a substitution matrix enriched by structural information. Two other datasets are issued from ProSup[34] which structurally aligns protein pairs and proposes two classifications of the resulting alignments; one satisfies the smallest RMSD and the other satisfies the largest number of equivalents. For each selected pair, both alignments have been considered, and for each alignment, only equivalent residues have been taken into account. The three datasets of structural alignments are called S_{HOM}, S_{RMSD}, and S_{EQ}, respectively. We count 427,579 amino acids pairs in S_{HOM}, 211, 807 in S_{RMSD}, 214, 525 in S_{EQ}.

Substitution Probability Tables for Distantly Related Protein Families

We constructed substitution probability tables for pairs (values in the table give the probability of a substitution of a residue at the top of a column by all other residues, with columns that sum to 1) of aa occurring within α-helices (HOM_α, $RMSD_\alpha$, EQ_α) and β-sheets (HOM_β, $RMSD_\beta$, EQ_β) in our three structurally aligned protein families. We compared them to substitution probability tables for α-helices (O_α) and β-sheets (O_β) proposed by Overington[13] and constructed from families of proteins in a range 25–80% of homology. We obtained 58%, and 64.6% correlation between HOM_α, O_α and HOM_β, O_β respectively. Similar values are obtained for S_{RMSD}, S_{EQ}. Note that there is 80.3%, 92.3% correlation between HOM_α, HOM_β and O_α, O_β. The low correlation found among our matrices and the Overington's matrices indicates the importance of constructing matrices from protein families with low homology. The high correlation of the matrices computed for α-helices (HOM_α, O_α), and similarly for β-sheets (HOM_β, O_β), is an indicator of a limited difference among aa substitution rates in α-helices and β-sheets. This supports the usage of structural blocks to align sequences which do not differentiate between α-helices and β-sheets.

Criteria to Compute Scores for Two Substitution Matrices

Substitution matrices are symmetric score matrices that represent the probability of pairs of aa i,j to mutate one in the other along evolution. To construct our matrices we followed the approach proposed for Blosum matrices $[s_{ij}]$[10]: a random substitution of i with j corresponds to the expected frequency of substitution f_{exp} and it has score 0; a favored substitution f (that is $f > f_{\text{exp}}$) has a positive score and a disadvantaged substitution f (that is $f < f_{\text{exp}}$) has a negative score. Scores $s_{i,j}$ are computed as follows:

$$\log_2\left(\frac{f_{\text{obs}_{ij}}}{f_{\text{exp}_{ij}}}\right) \quad \text{with} \quad f_{\text{exp}_{ij}} = \begin{cases} f_{\text{obs}_i} f_{\text{obs}_j} & i = j \\ 2 \cdot f_{\text{obs}_i} f_{\text{obs}_j} & i \neq j \end{cases}$$

where $f_{\text{obs}_{ij}}$ is the observed frequency substitution of residues i and j. To adapt the original score formula to pattern evolution, we tested three modified versions of the original score definition given above. The first version considers only residues present in a pattern type, and relates the frequency of the pair i,j observed in the pattern $\left(f^{\text{pat}}_{obs_{ij}}\right)$ to the expected frequency $\left(f^{\text{pat}}_{exp_{ij}}\right)$

$$\log_2\left(\frac{f^{\text{pat}}_{\text{obs}_{ij}}}{f^{\text{pat}}_{\text{exp}_{ij}}}\right) \tag{1}$$

where, for the overlapping hypothesis, $f^{\text{pat}}_{exp_{ij}}$, is defined as $c \times f^{\text{pat}}_{obs_i} \times f^{\text{pat}}_{obs_j} + f^{\text{outpat}}_{obs_i} \times f^{\text{pat}}_{obs_j} + f^{\text{pat}}_{obs_i} \times f^{\text{outpat}}_{obs_j}$ with

$f_{obs_i}^{\text{outpat}}$ denoting the frequency of i calculated outside patterns, and for the nonoverlapping hypothesis as $c \times f_{obs_i}^{\text{pat}} \times f_{obs_j}^{\text{pat}}$, with $c = 2$ if $i \neq j$, and $c = 1$ otherwise.

The second proposition introduces a weight $\omega_{i,j}$ in Eq. (1) to calibrate pairs distribution. This weight is the ratio between the number of pairs ij observed in patterns and the number of pairs ij observed in the whole sequence

$$\log_2\left(\frac{f_{\text{obs}_{ij}}^{\text{pat}}}{f_{\text{exp}_{ij}}^{\text{pat}}} \times \omega_{ij}\right), \quad \text{with} \quad \omega_{ij} = \frac{f_{\text{obs}_{ij}}^{\text{pat}}}{f_{\text{obs}_{ij}}^{\text{seq}}} \tag{2}$$

where $f_{exp_{ij}}^{\text{pat}}$ is defined as above.

The third proposition considers the observed frequency of ij pairs in patterns and compares it with the expected frequency of ij in the sequence ($f_{exp_{ij}}^{\text{seq}}$)

$$\log_2\left(\frac{f_{\text{obs}_{ij}}^{\text{pat}}}{f_{\text{exp}_{ij}}^{\text{seq}}}\right) \tag{3}$$

where $f_{exp_{ij}}^{\text{seq}}$ is defined as $f_{exp_{ij}}$ in the original score.

Corresponding versions of these three propositions, where frequencies of pairs sitting outside patterns are considered instead, allow to construct substitution matrices for aa occurring outside a pattern type. Matrices are multiplied by a scaling factor of 3 and then rounded to the nearest integer value.

The Algorithm

The combinatorial information relative to blocks of hydrophobic aa is handled at the algorithmic level during alignment by using two substitution matrices and gap penalties which are dependent on hb overlapping. PHYBAL analysis is based on two steps: a 1D screening detecting hb for all input sequences, and a pairwise alignment algorithm applied to these sequences and their combinatorial structure of hb.

The basic alignment method

The alignment is based on the dynamic programming algorithm of Needleman and Wunsch[30] and two substitution matrices (described later), one used to compare pairs of residues where none of them belongs to a block, and the other to compare pairs of residues where at least one of them belongs to a block. This strategy in comparing residues corresponds to the *overlapping hypothesis* (see Fig. 2) and fits the statistical conclusions reached in selecting best matrices. Two kinds of gap penalties are also used, one for gap introduction in hb (gap opening is referred to as bGOP, gap extension as bGEP) and the other outside hb (GOP and GEP) leading to a 4D gap space. PHYBAL-2D indicates PHYBAL when run within the 2D space of gap values defined by GOP = bGOP and GEP = bGEP. A block remains a single entity even if gaps are inserted. End gap opening is set to 2 and end gap extension is set to 1, with the idea in mind not to penalize small insertion/deletion at the sequence extremes.

Global and global-local alignment

The program performs *global* alignment, which assumes that the two proteins are comparable over the entire length of one another, as well as *global–local* alignment, which assumes an overlap of the two sequences and does not penalize insertion/deletion at the end of the alignment (end gap opening = end gap extension = 0).

Score of a pairwise alignment

It is the sum of scores between pairs of letters in the columns of the alignment normalized by the length of the alignment: $S(A) = \frac{1}{N}\sum_{j=1}^{N} w(x_j^1, x_j^2)$, where $w(x_j^1, x_j^2)$ is either the value of the substitution of residue x_j^1 with residue x_j^2, or a (opening, extension) gap penalty at alignment position j, and N is the number of columns in the pairwise alignment.

Comparison of Various Substitution Matrices With IHBM and OHBM

PHYBAL and PHYBAL-2D running with IHBM and OHBM have been compared with PHYBAL-2D run with several substitution matrices: Gonnet,[9] Blosum30, Blosum45, Blosum62,[10] HSDM,[35] PAM250,[8] Johnson,[36] and Remote Homo,[37] where the same matrix models the alignment in and out blocks and the algorithm becomes Needleman and Wunsch algorithm.[30] Blosum30, HSDM, and Remote Homo have been constructed to align distantly related proteins. Johnson and HSDM matrices are originally floating point matrices but we use their discretization.

Three Reference Sets of Protein Pairs Used to Analyze and Validate PHYBAL

We define three datasets of protein pairs with <30% sequence identity to analyze PHYBAL performance.

Set of eight randomly selected protein pairs

The structural alignment programs MATRAS[38] and SSM[39,40] have been used to align eight pairs of proteins known to have low homology, varying within the range 5.6–26.4% of amino acids (aa) identity, covering a wide spectrum of lengths 60–380aa, and classified as α + β (1), all α (4), and all β (3). They are Methyl CpG binding domain (α-chain, 1D9N, *H. sapiens*, 92aa) and Methyl CpG binding domain 2 (α-chain, 1QK9, *H. sapiens*, 74aa) (pair P1); Hemoglobine (α-chain, 1HGA, *H. sapiens*, 141aa) and Myoglobine (1MBO, *P. catodon*, 153aa) (P2); Hemoglobine (β-chain, 1HGA, *H. sapiens*, 146aa) and Leghemoglobine (1GDI, *L. luteus*, 159aa) (P3); ribosomal protein L20 (α-chain, 1GYZ, *A. aeolicus*, 59aa) and Poly(A) binding protein (α-chain, 1I2T, *H. sapiens*, 60aa) (P4); C-phycocyanine (β-chain, 1CPC, *F. diplosiphon*,

172aa) and Myoglobine (1MBO, *P. catodon*, 153aa) (P5); Plastocyanine (1AG6, *S. oleracea*, 99aa) and Azurin (1AZU, *P. aeruginosa*, 121aa) (P6); Tick-Borne Encephalitis (TBE) virus capside protein (1SVB, 395aa) and Semliki Forest virus (SFV) capside protein (α-chain 1RER, 383aa) (P7); V8 protease (α-chain, 1QY6, *S. aureus*, 216aa) and Trypsine (1SGT, *S. griseus*, 223aa) (P8). Because of the different behavior of MATRAS and SSM on our reference set, alignments have been verified and modified by hand when necessary, and we defined a Structural Alignment Difficulty (SAD-)coefficient for each pair of proteins to be $1-M/N$, where M is the number of amino-acid pairs shared by the two alignments and N is the average length of the two alignments.

The Domingues dataset

The 127 pairs of proteins constituting the Domingues dataset[41] present <30% sequence identity, at least 35 equivalent residues and share all secondary structural elements in the hydrophobic core. It provides a complete list of structural alignments, with RMS errors <3 Å, produced with PROSUP (possibly several ones for the same pair of structures). See Supplementary Table III. Comparison of predicted alignments (realized with PHYBAL and PHYBAL-2D) to Domingues reference alignments only takes into account *equivalent* residues, i.e., residues which are ≤5 Å apart. For each predicted alignment, we consider the best CAP obtained from the comparison to all alternatives in the dataset (as suggested in Ref. 41).

Set issued from BAliBASE 2.0

Out of reference set 1 in BAliBASE 2.0,[42] we considered all pairwise alignments of <25% sequence identity extracted from multiple alignments. We obtained 181 pairs of sequences and divided them into six different levels of percentage identity: ≤12% (12), 12–15% (i.e., >12% and ≤15%) (40), 15–17% (29), 17–20% (41), 20–22% (18), and 22–25% (41). See Supplementary Table IV. Comparison of predicted alignments (realized with PHYBAL, PHYBAL-2D, and HMMSUM) to BAliBASE 2.0 reference alignments only takes into account residues within *core blocks*, i.e., regions that can be reliably aligned.

Overlap of the datasets

The overlap is estimated by comparing pdb names of proteins in the datasets. HOMSTRAD, BAliBASE 2.0 and Domingues datasets contain 523, 124, and 165 proteins, respectively. Proteins shared by the datasets are as follows: 25 for HOMSTRAD-Domingues, 17 for HOMSTRAD-BAliBASE, and 12 for Domingues-BAliBASE. No protein within the eight selected pairs belongs to the datasets.

HMMSUM on BAliBASE 2.0

Five pairs of sequences (1, 3, 1 for tests ≤12%, 12–15%, 15–17%, respectively) for BAliBASE 2.0 use the X character to represent unidentified aa. Since HMMSUM does not handle the X character, we deleted the sequences from the corresponding dataset used to compare PHYBAL and HMMSUM.

Solvent Accessibility

Residues surface solvent accessibility has been calculated with NACCESS 2.1.1[43] with a probe size of 1.4 Å. As indicated in the NACCESS reference manual, relative accessibilities are calculated for each aa in a protein by expressing the summed residue accessible surfaces as a percentage of that observed in a ALA-X-ALA tripeptide. Tripeptides are built using the QUANTA molecular graphics package in extended conformations, so as to expose the central X residue in the tripeptide as much as would normally be possible in a protein. Because of unusual bond angles, bond lengths and distorted geometry in real proteins, these values can often exceed 100% (as seen in the x-coordinates in Fig. 1, bottom). A residue with ≤30% accessibility is considered as buried.

RESULTS

Properties of Hydrophobic Blocks, Secondary Structures, and Solvent Accessibility

Hb are highly correlated with rss and low solvent accessibility: 85.74% of detected hb on the 613 sequences coming from the HOMSTRAD database, share at least one residue with a rss, and 89.83% of rss share at least one residue with a detected hb (86.94% for α-helices and 92.42% for β-sheets). Hb sharing at least one residue with a rss are called *true hydrophobic blocks* (thb), and hb which are not thb are called fhb (f stands for false). The 70.72% and 76.36% of residues sitting in hb and thb overlap rss; the 67.28% of residues sitting in rss overlap hb. The average length of rss, thb, and fhb is 8.1aa (with $\sigma = 5.6$), 8.8aa (with $\sigma = 6.1$), and 4.2aa (with $\sigma = 2.6$) (see Fig. 1, top). The average relative solvent accessibility surface for rss, thb, fhb is 25.3% (with $\sigma = 24.8$), 23.7% (with $\sigma = 24.7$), 33.2% (with $\sigma = 28.6$) per residue (see Fig. 1, center). In particular, the average relative solvent accessibility surface for residues sitting outside rss is 40.5% (with $\sigma = 30.4$), and for residues sitting outside hb is 40.6% (with $\sigma = 29.35$) (see Fig. 1, bottom). These values suggest that we can meaningfully distinguish two kinds of hb: thb which are closer to rss and probably involved in the folding stability, and fhb that we can consider either as false positives or as blocks undergone a different evolutionary pressure than thb, and involved in molecular interaction (as indicated above, fhb tend to appear on the protein surface rather than in its interior).

Conservation Properties of Structural Patterns

We consider five different structural environments where to analyze residue evolution: rss, hb, thb, regions defined by the overlapping between rss and thb

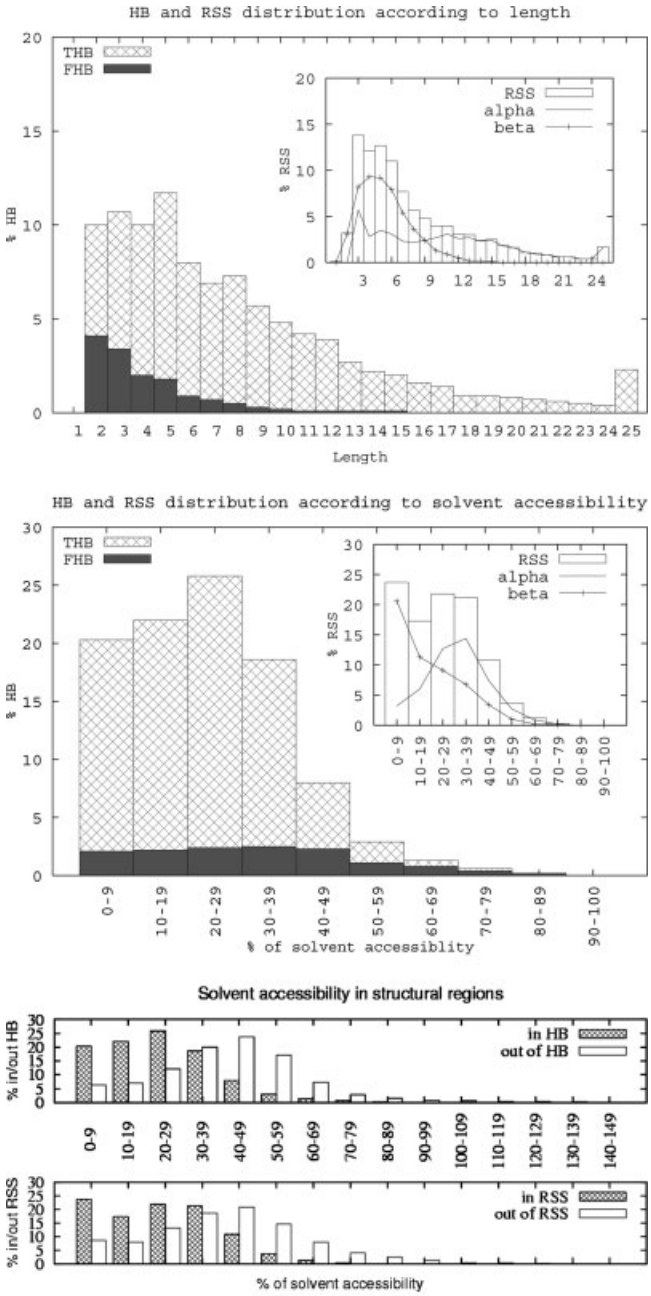

Fig. 1. Top: profile of hb distribution with respect to protein length for our 613 protein sequences. Percentage of hb (*y*-axis) whose length (intended to be the number of aa, *x*-axis) is plotted with a composite bar graph representing thb and fhb proportions. Profiles for rss are given on top right of bar graphs. Center: profile of hb distribution with respect to solvent accessibility. Percentage of hb (*y*-axis) whose average residue surface accessibility falls within a given range (*x*-axis) is plotted. Bottom: solvent accessibility profile in and out hb/rss plotted with grey and white bars. (Details on solvent accessibility values in Materials and Methods.)

(rss&thb), and br. Any of these five different environments is referred to as a *pattern*. Patterns have been considered with the idea in mind to determine the most suitable structures for the treatment of weak homologies

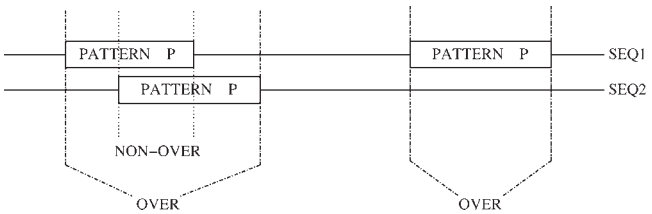

Fig. 2. Two sequences with patterns of type P (i.e., hb, rss, thb, hb&rss, br) displaying two distinct readings of overlapping patterns: NON-OVER where residues occur in a pattern for both sequences, OVER where at least one residue belongs to a pattern.

and for describing evolutionary pressure inside and outside patterns.

Pattern alignment and the overlapping condition

The alignment of two sequences might present overlapping of a given pattern type, as illustrated in Figure 2. Since the extremes of a pattern in a sequence might overlap with nonstructural residues of the other sequence because of size difference, we can explicitly consider overlapping nonstructural regions to form a pattern or not (Fig. 2). In the first case we speak of *overlapping hypothesis* and consider certain nonstructural residues as prone to form a pattern, hence subjected to the same evolutionary pressure as residues lying in patterns. In the second case we speak of *nonoverlapping hypothesis* and consider nonstructural residues as resulting from weaker evolutionary pressures and more likely to be randomly distributed. The 44% of residues sitting in a hb, is aligned with a residue which is also sitting within a hb. The 56% holds for rss. These values support the interest in analyzing the two hypothesis above.

Properties of pairs of residues lying within patterns of structurally aligned sequences

We analyzed conservation of acidity, basicity, ionizability, aromaticity, polarity, apolarity, and hydrophobicity, together with aa volume conservation within sequences and patterns. Physico-chemical properties appear more conserved in patterns than in sequences, as well as sequence identity. The most conserved properties are hydrophobicity and aromaticity, where three residues over the four characterizing this latter are hydrophobic. High conservation among large residues (with volume 185–230 $Å^3$) which are all hydrophobic shows that hydrophobic residues are likely to be important for structural reasons. This analysis is in agreement with the conclusion of Kinjo and Nishikawa.[44] See Supplementary Table I. (Values are given for S_{HOM}, and those obtained for S_{RMSD} and S_{EQ} are comparable; the three sets of structural alignments S_{HOM}, S_{RMSD}, and S_{EQ} are defined in Material and Methods.)

Selection of Two Best Fitting Matrices of Amino Acid Substitution in and out Hydrophobic Blocks

We generated 90 different matrices for aa substitution within patterns issued by the combination of the following

	A	R	N	D	B	C	Q	E	Z	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3	-1	-2	-3	-3	0	-1	-1	-1	-1	-2	0	0	-1	1	0	-5	0	-1	0	-1	1
R	-1	6	0	-1	-1	-4	1	0	0	-3	0	-2	-1	2	0	-2	-4	0	-1	-1	0	-2
N	-2	0	5	1	1	-3	0	0	0	-1	0	-3	-2	0	-1	-2	-3	0	0	-2	-1	-2
D	-3	-1	1	6	6	-5	0	1	1	-1	-2	-4	-3	0	-3	-3	-3	0	-1	-1	-1	-3
B	-3	-1	1	6	6	-5	0	1	1	-1	-2	-4	-3	0	-3	-3	-3	0	-1	-1	-1	-3
C	0	-4	-3	-5	-5	10	-3	-4	-4	-2	-4	1	1	-4	0	0	-8	-1	-2	-1	0	2
Q	-1	1	0	0	0	-3	4	2	2	-2	0	-2	-1	1	0	-1	-4	0	-1	0	0	-1
E	-1	0	0	1	1	-4	2	4	4	-3	-1	-3	-3	1	-2	-3	-4	0	0	-1	-1	-2
Z	-1	0	0	1	1	-4	2	4	4	-3	-1	-3	-3	1	-2	-3	-4	0	0	-1	-1	-2
G	-1	-3	-1	-1	-1	-2	-2	-3	-3	6	-2	-3	-3	-2	-2	-3	-5	0	-2	-3	-2	-2
H	-2	0	0	-2	-2	-4	0	-1	-1	-2	7	-2	-1	0	-1	0	-4	-1	-1	1	1	-2
I	0	-2	-3	-4	-4	1	-2	-3	-3	-3	-2	6	4	-2	3	2	-2	-2	0	0	1	5
L	0	-1	-2	-3	-3	1	-1	-3	-3	-3	-1	4	6	-2	4	3	-2	-1	0	1	1	3
K	-1	2	0	0	0	-4	1	1	1	-2	0	-2	-2	4	0	-2	-3	0	-1	-2	-1	-2
M	1	0	-1	-3	-3	0	0	-2	-2	-2	-1	3	4	0	7	2	-2	0	0	1	2	2
F	0	-2	-2	-3	-3	0	-1	-3	-3	-3	0	2	3	-2	2	8	-3	-2	-1	4	5	2
P	-5	-4	-3	-3	-3	-8	-4	-4	-4	-5	-4	-2	-2	-3	-2	-3	-13	-3	-4	-2	-2	-2
S	0	0	0	0	0	-1	0	0	0	0	-1	-2	-1	0	0	-2	-3	3	1	-1	-1	-1
T	-1	-1	0	-1	-1	-2	-1	0	0	-2	-1	0	0	-1	0	-1	-4	1	3	-1	0	1
W	0	-1	-2	-1	-1	-1	0	-1	-1	-3	1	0	1	-2	1	4	-2	-1	-1	13	5	0
Y	-1	0	-1	-1	-1	0	0	-1	-1	-2	1	1	1	-1	2	5	-2	-1	0	5	8	0
V	1	-2	-2	-3	-3	2	-1	-2	-2	-2	-2	5	3	-2	2	2	-2	-1	1	0	0	6

	A	R	N	D	B	C	Q	E	Z	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	6	1	1	1	1	3	1	1	1	2	1	-8	-8	1	-7	-7	4	3	2	-5	-8	-6
R	1	8	2	1	1	-2	3	2	2	0	2	-9	-8	4	-7	-6	2	1	1	-7	-5	-8
N	1	2	9	4	4	0	3	1	1	3	4	-10	-9	2	-9	-5	3	4	3	-6	-4	-8
D	1	1	4	9	9	-1	3	4	4	1	1	-10	-9	1	-9	-8	3	2	2	-6	-6	-9
B	1	1	4	9	9	-1	3	4	4	1	1	-10	-9	1	-9	-8	3	2	2	-6	-6	-9
C	3	-2	0	-1	-1	16	0	-2	-2	0	1	-9	-7	-2	-7	-7	2	2	1	-9	-8	-9
Q	1	3	3	3	3	0	6	4	4	1	3	-7	-7	3	-5	-7	3	2	2	-4	-4	-7
E	1	2	1	4	4	-2	4	7	7	1	2	-11	-9	3	-7	-9	4	2	2	-7	-8	-8
Z	1	2	1	4	4	-2	4	7	7	1	2	-11	-9	3	-7	-9	4	2	2	-7	-8	-8
G	2	0	3	1	1	0	1	1	1	10	1	-11	-8	1	-6	-7	3	3	1	-6	-5	-8
H	1	2	4	1	1	1	3	2	2	1	13	-8	-8	1	-6	-5	2	2	1	-4	-4	-6
I	-8	-9	-10	-10	-10	-9	-7	-11	-11	-11	-8	-4	-7	-9	-8	-8	-4	-10	-5	-9	-8	-6
L	-8	-8	-9	-9	-9	-7	-7	-9	-9	-8	-8	-7	-3	-8	-4	-7	-4	-8	-6	-10	-8	-7
K	1	4	2	1	1	-2	3	3	3	1	1	-9	-8	6	-8	-9	4	2	2	-6	-7	-8
M	-7	-7	-9	-9	-9	-7	-5	-7	-7	-6	-6	-8	-4	-8	-1	-6	-4	-8	-3	-4	-7	-9
F	-7	-6	-5	-8	-8	-7	-7	-9	-9	-7	-5	-8	-7	-9	-6	1	-4	-6	-5	-4	-3	-10
P	4	2	3	3	3	2	3	4	4	3	2	-4	-4	4	-4	-4	15	4	3	-2	-4	-2
S	3	1	4	2	2	2	2	2	2	3	2	-10	-8	2	-8	-6	4	6	4	-4	-6	-7
T	2	1	3	2	2	1	2	2	2	1	1	-5	-6	2	-3	-5	3	4	7	-4	-5	-5
W	-5	-7	-6	-6	-6	-9	-4	-7	-7	-6	-4	-9	-10	-6	-4	-4	-2	-4	-4	3	-3	-9
Y	-8	-5	-4	-6	-6	-8	-4	-8	-8	-5	-4	-8	-8	-7	-7	-3	-4	-6	-5	-3	3	-11
V	-6	-8	-8	-9	-9	-9	-7	-8	-8	-8	-6	-6	-7	-8	-9	-10	-2	-7	-5	-9	-11	-6

Fig. 3. Top: Inside Hydrophobic Blocks Matrix IHBM. Bottom: Outside Hydrophobic Blocks Matrix OHBM.

TABLE I. Best PCAP and Corresponding CAP (in Parenthesis) Obtained for Alignments of Protein Pairs *P*1–*P*8 with Different Alignment Methods (Optimized Gap Values are Determined for Each Pair) and GEP/bGEP Varying from 0 to 10, GOP/bGOP Varying from 0 to 20

	Dataset of eight pairs of proteins							
Method	P1	P2	P3	P4	P5	P6	P7	P8
PHYBAL[a]	71.6 (53)	**96.8 (138)**	**89.7 (132)**	**77.3 (42)**	34.8 (29)	**62.5** (58)	**35.1 (124)**	48.8 (94)
PHYBAL-2D[b] + IHBM&OHBM	67.0 (48)	92.2 (134)	**89.7 (132)**	70.8 (38)	22.9 (18)	55.0 (52)	26.8 (91)	46.4 (94)
Gonnet	**75.0 (56)**	91.5 (133)	88.5 (131)	70.8 (38)	37.0 (33)	49.6 (49)	29.2 (99)	46.1 (94)
HSDM	68.8 (50)	90.8 (131)	**89.7 (132)**	59.1 (31)	24.4 (9)	58.1 (57)	21.8 (74)	50.4 (94)
Blosum 62	**75.0 (56)**	91.5 (133)	83.3 (123)	70.8 (38)	**41.4 (40)**	56.6 (54)	27.7 (92)	41.7 (78)
Blosum 45	**75.0 (56)**	91.5 (133)	85.3 (127)	70.8 (38)	**41.4 (40)**	57.3 (58)	16.2 (34)	41.8 (79)
Blosum 30	**75.0 (56)**	93.5 (134)	72.8 (109)	70.8 (39)	35.4 (39)	49.2 (52)	19.7 (52)	39.5 (72)
PAM250	**75.0 (56)**	96.1 (137)	88.5 (131)	46.3 (27)	34.8 (29)	46.8 (48)	25.4 (86)	42.0 (83)
Johnson	60.4 (46)	94.8 (137)	88.5 (131)	25.0 (11)	34.6 (29)	54.7 (50)	18.3 (55)	49.2 (94)
Remote Homo	63.8 (48)	89.6 (128)	79.4 (118)	50.7 (28)	34.6 (29)	50.8 (50)	15.9 (46)	**59.1 (111)**
SAD-coefficient[c]	0.39	0.06	0.31	0.33	0.26	0.31	0.59	0.1
RMSD[c]	2.53	2.48	1.47	2.41	2.70	2.44	3.42	1.57
% identity[c]	24	26	16	13	5	9	9	11

Bold characters represent best performance.
[a]PHYBAL best PCAP and CAP values.
[c]PHYBAL-2D best PCAP and CAP are calculated on different matrices.
[c]Pair divergence in sequence (% identity) and structure (SAD-coefficient and RMSD).

four parameters: (i) a pattern (rss, thb&rss, thb, hb, br), (ii) a score equation [(1), (2), (3)], (iii) overlapping or nonoverlapping hypothesis, (iv) a dataset of structural alignments (S_{HOM}, S_{RMSD}, S_{EQ}). A complementary pool of 90 matrices has been calculated for aa substitution outside patterns. For each one of the 90 combinations of parameters, we tested the behavior of the corresponding pair of matrices describing substitution rates within and without patterns using PHYBAL-2D on our reference set constituted by eight pairs of proteins, with all gap values varying from 0 to 10 for extension and from 0 to 20 for opening.

Matrices behavior is analyzed with respect to average best Percentage of Correctly Aligned Pairs (PCAP), the number of Correctly Aligned Pairs (CAP), stability and gap coherence. *PCAP* is the number of correctly aligned pairs (of aa, or gap and aa) calculated on the structural alignment over the length of the predicted alignment. PCAP variability is graphically represented by colored matrices of 11 × 21 entries (corresponding to gap combinations), referred to as PCAP landscapes in Figure 4 and Supplementary Figure 1. *CAP* is the number of correctly aligned pairs of aa calculated on the structural alignment. *Stability* measures the robustness of the system to variability of gap values and corresponds to the size of monochromatic regions associated to high PCAPs in the PCAP landscape. *Gap coherence* tests whether best PCAPs are obtained on combinations where gap opening is larger than gap extension, and this corresponds to verify that best PCAPs lie above the PCAP landscape diagonal GOP = GEP (and bGOP = bGEP when a 4D gap space is considered). This property ensures that best PCAPs are not reached with lots of small insertions, which is intuitively expected for GEP > GOP (bGEP > bGOP). From the analysis of the 90 pairs of matrices we concluded that (1) Eq. (2) does not provide competitive PCAPs, (2) high PCAPs and good stability are obtained with Eq. (3) on both the overlapping and nonoverlapping hypothesis, (3) PCAP and stability calculated for S_{HOM} are worse than for S_{RMSD} and S_{EQ}, (4) all PCAP landscapes associated to Eqs. (1) and (3) present gap coherence. Best PCAP (54.5%), high CAP (548 over 1196 predicted by PHYBAL-2D, and over 1152 after structural alignment) and high stability (31 over 231 contiguous gap combinations giving >50% PCAP) is determined for (i) thb, (ii) Eq. (3), (iii) overlapping hypothesis, (iv) S_{EQ}. The two matrices associated to these four conditions have been selected and their average PCAP landscape is reported on the top right of Figure 4. Similar properties are obtained for other sets of conditions: thb, Eq. (3), overlapping hypothesis, and S_{RMSD}; rss, Eq. (3), nonoverlapping hypothesis, and S_{RMSD} or S_{EQ}. (See their PCAP landscapes in Supplementary Fig. 1, together with the full account of the analysis.)

The selected pair of matrices (displayed in Fig. 3) are called Inside Hydrophobic Blocks Matrix (IHBM) and Outside Hydrophobic Blocks Matrix (OHBM). In OHBM, hydrophobic amino acid pairs attain negative scores and this holds, in some extent, for hydrophobic aa identities also. A high identity score is found on proline, cysteine, histidine, and glycine. On the contrary, in IHBM, substitution scores of hydrophobic residues increase and hydrophobic aa identities are all positive. Proline identity is negative since no proline belongs to hb.

Comparison of IHBM and OHBM With Other Matrices on the Eight Selected Protein Pairs

PHYBAL-2D and PHYBAL have been run on the selected set of eight protein pairs with substitution matrices

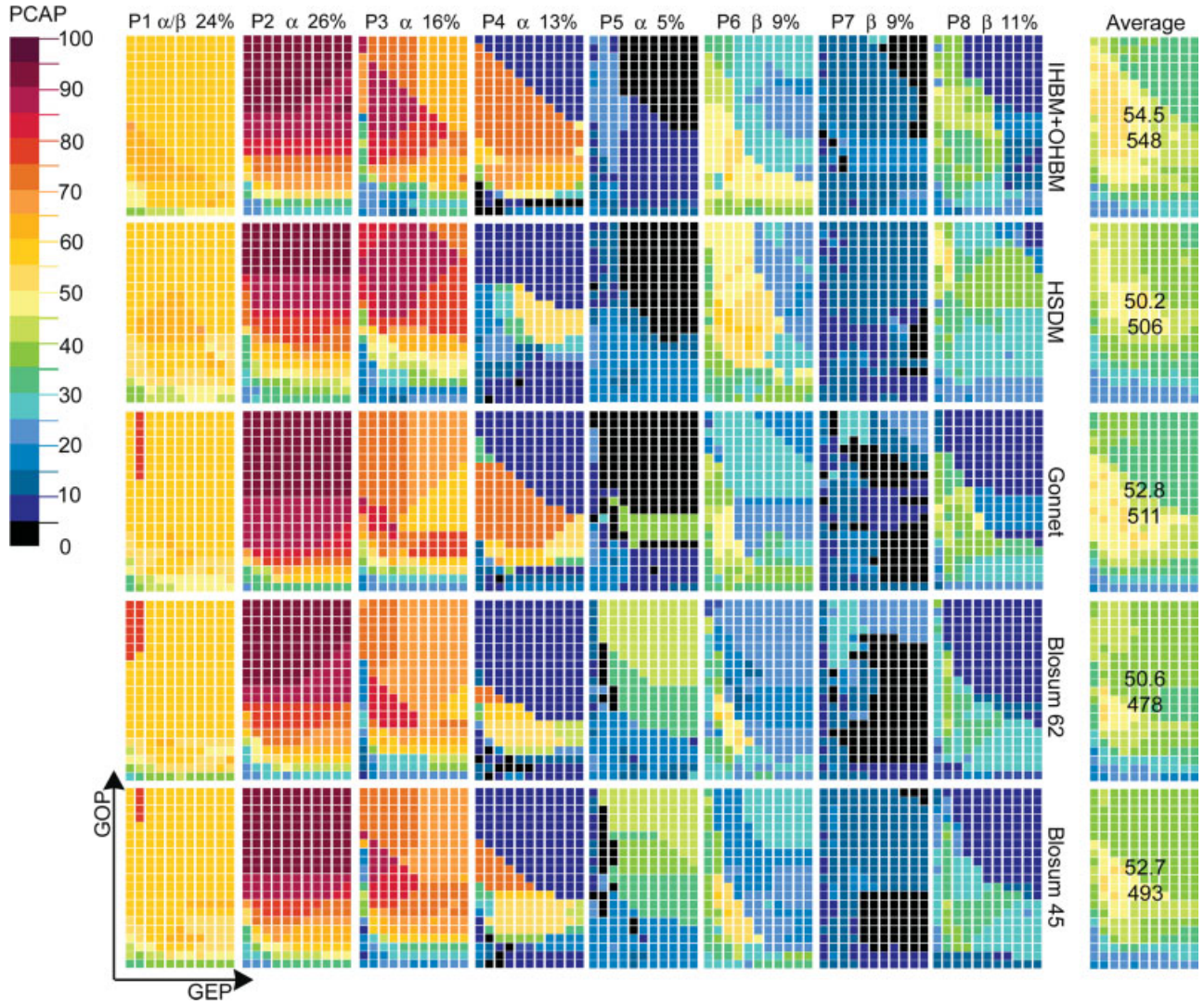

Fig. 4. Left: Alignment performance of PHYBAL-2D obtained on protein pairs *P*1-*P*8 (structural classes and percentage of identity are indicated on column headings) with IHBM and OHBM, HSDM, Gonnet, Blosum 62, Blosum 45. Blosum 30, Johnson, PAM250, and Remote Homo are omitted because reaching lower average PCAPs. Alignments corresponding to GEP values going from 0 to 10 and GOP values going from 0 to 20 are represented on a 11 × 21 2D PCAP landscape (with GEP on *x*-axis and GOP on *y*-axis): each alignment has been compared with the corresponding structural alignment and its PCAP is indicated by a suitable color. Right: 2D PCAP landscapes of average PCAPs computed over all eight protein pairs are represented for each alignment method. Best average PCAPs (top) and CAPs (bottom) are reported.

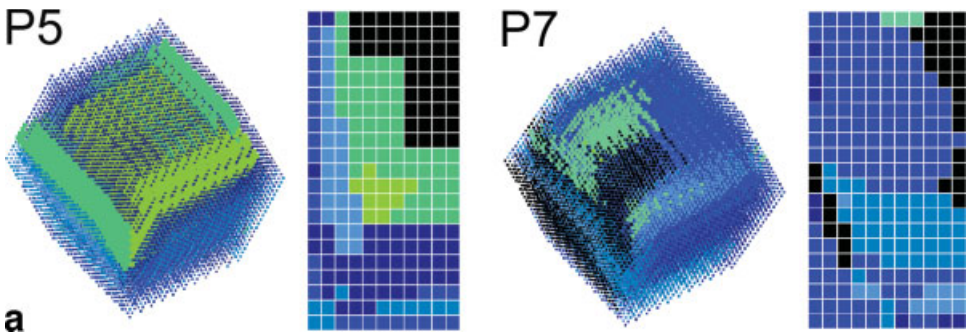

Fig. 5. (a) PCAP landscapes of P5 and P7 obtained with PHYBAL-2D (2D) and PHYBAL [(3D, after Principal Component Analysis (PCA)] based on a global-local alignment algorithm. Compare with columns P5 and P7 in Figure 4.

Blosum 30, Blosum 45, Blosum 62, HSDM, Gonnet, Johnson, and Remote Homo besides IHBM & OHBM. Substitutions in and out hb are treated by the same matrix. Comparison in the 2D gap space leads to observe that using suitable matrices according to sequence specificity helps to improve the alignment,[15] and comparison in the 4D gap space confirms that considering structural information in treating gap weights[14] improves alignments further. The wide spectrum of structural properties covered by the eight protein pairs is instructive to appreciate these two hypothesis. Validation of them is realized below on two large sets of proteins.

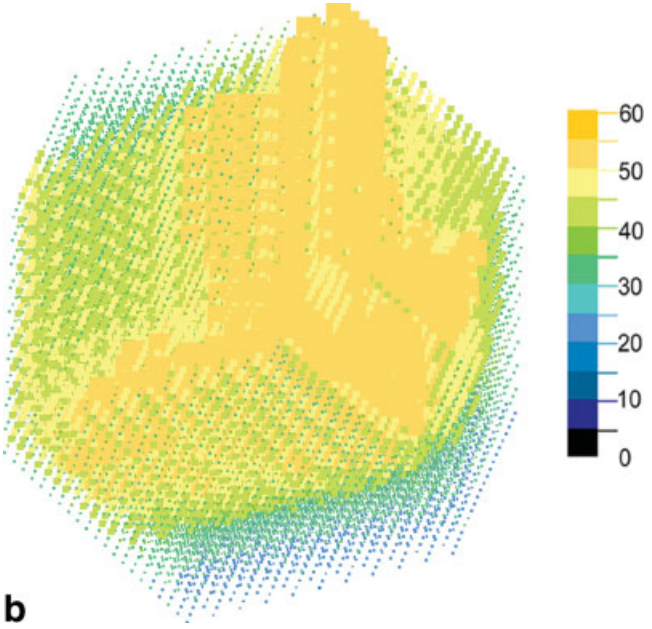

Fig. 5 (Continued). (b) PCAP landscape of average PCAPs obtained with PHYBAL on the eight protein pairs (after PCA). The yellow area represents the hot-spot region where PHYBAL performs the best (color scale as in Fig. 4).

Analysis in two dimensions

For each matrix, we output 231 alignments, corresponding to the 11 × 21 combinations of gap values. Each alignment is compared with the corresponding structural alignment and PCAP is calculated. Results are reported in Table I and Figure 4. For the eight pairs of proteins, 54.5% average PCAP is obtained for PHYBAL-2D with IHBM&OHBM with a large region of 31 gap combinations reaching 50–55% PCAP and localized between GOP of 7–17 and GEP of 0–3, thus displaying the best PCAP, stability and gap coherence (Fig. 4, right). Gonnet presents good stability for 45–50% PCAP but an unstable behavior for best PCAPs (50–55%). Much less stability is shown by the other matrices. In particular, HSDM presents stable regions on (almost) each protein pair which are localized on different areas of the PCAP landscapes, leading to a low average stability suggesting a high sensibility of the system to gap variations (see Fig. 4).

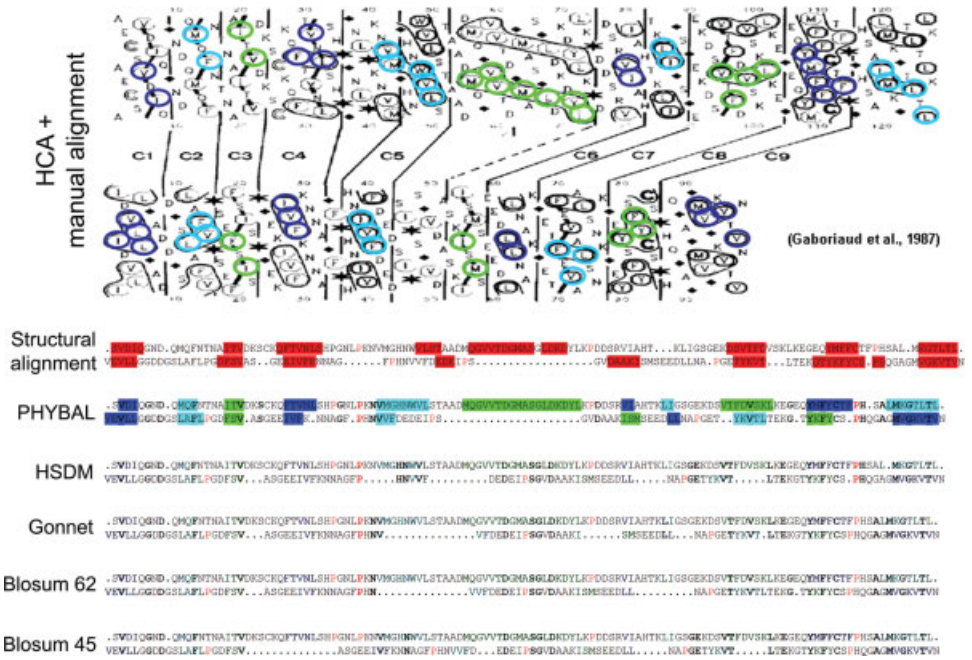

Fig. 6. Alignments of protein pair P6 (Plastocyanin and Azurin, β, 13% identity), obtained manually after HCA analysis (first row), and automatically with PHYBAL (row 3), and with PHYBAL-2D on several matrices (rows 4–7). Rss are highlighted in red on structural alignment (row 2). Hydrophobic clusters are highlighted in the HCA plot by using the same colors highlighting corresponding hb in PHYBAL. Notice the central region of the structural alignment where a large gap is inserted; PHYBAL recognizes the gap insertion which is missed by PHYBAL-2D run on all other matrices.

TABLE II. Best average PCAP and Corresponding CAP (in Parenthesis) Obtained for Alignments of Protein Pairs in BAliBASE (A) and Domingues (B) Datasets with Different Alignment Methods (Optimized Gap Values are Determined for Each Pair)

(A)

Method	≤12	glo	loc	12–15	glo	loc	15–17	glo	loc	17–20	glo	loc	20–22	glo	loc	22–25	glo	loc
PHYBAL hot-spot	**43.2 (296)**	10	3	**39.0 (1203)**	11	3	**40.1 (1289)**	80	21	**52.8 (2189)**	161	41	57.9 (1054)	0	80	75.0 (3737)	120	42
PHYBAL-2D + IHBM &OHBM	41.6 (244)	9	4	37.3 (1018)	5	4	37.7 (1211)	9	4	49.4 (1971)	11	13	57.0 (1048)	0	16	72.1 (3614)	10	16
Gonnet	37.8 (246)	0	3	29.8 (816)	0	20	35.5 (1145)	8	2	51.9 (2025)	10	6	55.5 (1094)	0	2	73.3 (3750)	19	16
HSDM	24.4 (176)	0	4	28.2 (823)	0	7	29.5 (1100)	0	5	48.6 (1934)	1	5	54.8 (986)	0	10	**75.9 (3813)**	21	3
Blosum 62	33.0 (239)	0	1	31.2 (964)	0	6	34.0 (1260)	0	12	49.8 (1993)	7	7	**63.4 (1279)**	2	1	74.9 (3852)	18	10
Blosum 45	36.3 (243)	0	7	31.4 (898)	0	3	33.7 (1207)	0	9	49.1 (1917)	8	10	57.1 (1087)	0	6	72.6 (3702)	8	10
Blosum 30	26.8 (197)	0	3	28.4 (830)	0	4	32.0 (1179)	0	5	45.3 (1884)	0	17	54.5 (856)	0	2	71.7 (3676)	7	11
PAM250	22.4 (174)	0	9	26.3 (821)	0	20	27.4 (905)	0	8	42.4 (1629)	0	3	54.0 (1101)	0	2	71.0 (3604)	1	8
Johnson	24.8 (148)	0	14	27.8 (809)	0	7	29.1 (927)	0	29	46.4 (1916)	0	1	55.8 (1092)	0	2	72.0 (3670)	6	8
Remote Homo	36.1 (233)	0	1	23.4 (658)	0	5	31.9 (965)	0	8	38.5 (1626)	0	7	48.7 (1042)	0	4	66.2 (3372)	0	14

(B)

Method	PCAP (CAP)	glo	loc
PHYBAL hot-spot	**43.7 (6461)**	324	110
PHYBAL-2D + IHBM&OHBM	42.6 (6268)	49	21
Gonnet	40.2 (5982)	17	22
HSDM	43.5 (6454)	76	21
Blosum 62	41.4 (6172)	27	19
Blosum 45	41.6 (6259)	34	23
Blosum 30	38.1 (5694)	0	18
PAM250	39.3 (5797)	2	17
Johnson	41.5 (6051)	42	32
Remote Homo	32.5 (4783)	0	61

Bold characters represent the best PCAP over all methods. Alignment local stability is represented by the number of gap combinations providing a PCAP greater than "average PCAP - 2%" (loc). Alignment global stability is represented by the number of gap combinations providing a PCAP greater than "best PCAP - 5%" (glo). PHYBAL is run on hot-spot (324 combinations) and PHYBAL-2D on the full 2D gap space (231 combinations).

Ranges $7 \leq GOP \leq 17$ and $0 \leq GEP \leq 3$ for PHYBAL-2D run with IHBM&OHBM confirm that modelling sequence alignment with expensive gap opening and cheaper gap extension is appropriate for distantly related proteins as for proteins with more than 30% homology.[45] The PCAP landscape region reaching best PCAP sharply corresponds to the optimal region previously computed for highly conserved sequences.[45] Gap coherence is less pronounced for PHYBAL-2D run with other matrices.

Analysis in four dimensions

PHYBAL was run by systematically varying the costs of gaps within the 4D space of coordinates GOP, bGOP, GEP, and bGEP, modelling variability of gap weights in and out hb. PHYBAL obtains considerably better alignments than those obtained by PHYBAL-2D as illustrated in Table I, A–B. Except for P3, improvements vary from 2.4% to 11.9%. No matrix successfully approaches PHYBAL performance on all pairs. In this respect, notice that best PCAPs for P1, P5, P8 are obtained on distinct matrices. The two most difficult cases treated are P5 and P7 which present relatively long insertions/deletions in sequence endings. CAP and PCAP of both pairs improved significantly in the 4D gap space: P7 has a SAD-coefficient of 0.59 demonstrating the difficulty for alignment, and yet PHYBAL can successfully obtain 35% PCAP; the same PCAP is obtained for P5 displaying only 5% of sequence identity. We applied PHYBAL based on a global-local alignment algorithm and ameliorate further the results with best PCAP of 36.8% for P5 and 35.6% for P7 (Fig. 5). More information on alignments of P5 and P7 are found in Supplementary Material.

PHYBAL Validation on BAliBASE and Domingues Datasets

We validated our computational approach to pairwise alignment of distantly related proteins on two large datasets of protein pairs. We introduced two new measures describing global and local stability of a system. *Global stability* measures an alignment system performance with respect to other alignment systems (for instance, PHYBAL against PHYBAL-2D run with all other matrices, or PHYBAL against HMMSUM) and it is tested by counting, for each alignment method, the

number of gap combinations giving a PCAP which is at most 5% away from the best PCAP obtained over all systems under evaluation. *Local stability* measures the robustness of a system by counting the number of gap combinations giving a PCAP which is at most 2% away from the highest obtained by the system itself. Strictly speaking, these two notions capture some weak form of stability (as defined for the selection of IHBM&OHBM) since they count the number of optimal gap combinations without requiring these gap combinations to form a *region* in the PCAP landscape; they are easy to compute and appear sufficient for system comparison. On the basis of the analysis of the eight pairs of proteins we defined a *hot-spot* of gap combinations which is supposed to capture stability and gap coherence of PHYBAL behavior. The hot-spot corresponds to 324 gap combinations (out of 53361 gap combinations for the full 4D gap space), after the intervals $10 \leq \text{GOP} \leq 15$, $1 \leq \text{GEP} \leq 3$, $13 \leq \text{bGOP} \leq 17$, $1 \leq \text{bGEP} \leq 4$, where $\text{bGOP} \geq \text{GOP}$. The size of the hot-spot makes possible the comparison between PHYBAL and PHYBAL-2D with 324 versus 231 gap combinations, corresponding to a constant factor 1.4.

Validation on Domingues dataset

PHYBAL obtains the best PCAP and best CAP on the Domingues dataset, as reported in Table II. In fact, it performs as well as HSDM, but notice that Domingues dataset has been used to construct the HSDM matrix[35] (out of 122 pairs of proteins selected for the matrix construction, 113 of them belong to the Domingues database) and an optimal performance of this matrix on the dataset is expected. PHYBAL obtains 324 combinations (that is all combinations defining the hot-spot) showing a very good global stability, compared with a few dozens combinations (over 231) obtained with other methods. It obtains 110 gap combinations with a PCAP close to the highest demonstrating alignment reproducibility of best PCAP under different parametrization.

Validation on BAliBASE 2.0 dataset

We run PHYBAL on the BAliBASE 2.0 dataset and consistently obtained best average PCAPs for all groups of proteins with <20% identity, with good global and local stability (Table II). Blosum 62 on group 20–22% and HSDM on group 22–25% perform best but with a very weak stability, followed by PHYBAL displaying a more stable behavior and best PCAPs on both groups with respect to all remaining matrices. Difference in performance between PHYBAL and PHYBAL-2D run with specific substitution matrices consistently increases with protein pair divergence.

Four-Dimensional Analysis of PHYBAL With Blosum62&Blosum62

We run PHYBAL with Blosum62&Blosum62 (in and out hb) on the BAliBASE 2.0 dataset. See Table III. Performance of PHYBAL-2D+Blosum62 is improved by calculations in the 4D space, but PHYBAL with IHBM&OHBM still obtains better results than with Blosum62&Blosum62 for sequences with <20% identity. This shows the importance of both hb fitting matrices and 4D gap space for aligning divergent proteins.

TABLE III. PHYBAL Run with IHBM&OHBM and BLOSUM62&BLOSUM62 in Full 4D Gap Space (53361 Combinations) and in Hot-Spot (324 Combinations) on the BAliBASE 2.0

Method	≤12	12–15	15–17	17–20	20–22	22–25
PHYBAL						
PCAP	**45.0**	**39.3**	**40.1**	**53.3**	59.9	75.1
CAP	277	1100	1289	2164	1084	3755
glo	570	477	1214	1003	0	282
loc	32	102	327	185	107	132
PHYBAL hot-spot						
PCAP	43.2	39.0	**40.1**	52.8	57.9	75.0
CAP	296	1203	1289	2189	1054	3737
glo	10	11	150	161	0	61
loc	3	3	21	4	80	42
PHYBAL+Blosum62&Blosum62						
PCAP	33.8	33.2	37.7	52.5	**65.7**	**77.6**
CAP	228	1032	1258	2095	1269	3901
glo	0	0	98	1030	58	1808
loc	67	273	26	40	3	114

Comparison Between PHYBAL and Other Alignment Methods

We compare PHYBAL with two alignment methods which use structural information extracted from sequences.

Comparison with HMMSUM

HMMSUM[18] is a local pairwise alignment method that does not require the structure of the protein to be known but predicts local structures using HMMSTR[46] and aligns pairs of sequences accordingly to 281 structural context-based aa substitution matrices. HMMSUM, as PHYBAL, is based on the idea that structural context in proteins contributes to the selective pressure and that context specific substitution matrices should help to better align. The comparison on the BAliBASE 2.0 dataset of the two methods is reported in Table IV. PHYBAL has been run on the hot-spot of 324 gap combinations and HMMSUM on the 2D gap space of 231 combinations using HMMSUM-D_{NS} model.

A fine analysis on the BAliBASE dataset shows that PHYBAL performs consistently better than HMMSUM on pairs with <17% sequence identity. For pairs >17% sequence identity, HMMSUM improves PHYBAL's PCAPs of at most 2%, and global and local stability of the two systems remain comparable. This might be explained by conservation signals which might be present over the whole sequence and that global alignment can capture and by an appropriate use of matrices and gap penalties better adapted to regions in the sequence subjected to different evolutionary pressure.

TABLE IV. PHYBAL and HMMSUM Performance are Compared on the Dataset Issued by BAliBASE 2.0 Through Best Average PCAP, CAP (Calculated for Best PCAP), Global and Local Stability Values

Method	≤12	12–15	15–17	17–20	20–22	22–25
PHYBAL hot-spot						
PCAP	**44.9**	**39.4**	**40.3**	52.8	57.9	75.0
CAP	**257**	**1037**	1231	**2189**	1054	3737
glo	17	18	78	93	167	74
loc	2	2	22	41	80	42
HMMSUM						
PCAP	38.8	35.1	38.0	**54.3**	**59.2**	**76.9**
CAP	200	927	**1278**	2062	**1138**	**3836**
glo	0	4	5	43	24	43
loc	12	8	5	16	14	20

PHYBAL has been run on hot-spot in the 4D gap space of 324 combinations. HMMSUM has been run with HMMSUM-D_{NS} on the 2-dimensional gap space of 231 combinations.

Comparison With Hydrophobic Cluster Analysis Manual Method

Previous studies based on HCA and manual alignment guided by hydrophobic clusters[47–49] demonstrated the power of using hydrophobic signals to align distantly related proteins. In Figure 6, we consider the Phycocyanin-Azurin protein pair (P6). The inadequate alignment provided by known substitution matrices motivated,[27] where the two proteins have been manually aligned after the analysis of hydrophobic clusters and phylogenetic relationship has been established. PHYBAL generates a similar alignment, but in an automatic way: it correctly detects the large insertion while all other matrices miss it, see Figure 6. (All alignments in Fig. 6 run with gap penalties that obtained best average PCAP on Domingues dataset; for PHYBAL, GOP = 10, GEP = 1, bGOP = 17, bGEP = 2).

DISCUSSION

The particularly biased distribution of hb residues in a protein sequence suggests that these residues might contribute crucial structural information, as for instance for the protein folding. In this respect, we can formulate two different hypothesis concerning their evolution depending on whether we consider any hb residues to be meaningful or only those occurring in rss. The first hypothesis suggests that hb might provide a predisposition for the sequence to form rss. In this case, hb residues would evolve accordingly together with a biased evolution of their local environment. The second hypothesis suggests that hb residues lying in rss might evolve differently than any other residue in the protein and that the existence of hydrophobic aa outside a rss could be justified by the hydrophobic character of the protein core. The use of hb for aligning distantly related proteins has allowed an improvement in alignment accuracy and stability supporting the idea that hb play a fundamental role in structural conservation during protein evolution.

Extension to Multiple Alignment and Detection of Remote Homologues

PHYBAL aligns pairs of aa sequences but the tool has been designed to easily allow for an extension to progressive multiple alignment. Improvements at the early stage of pairwise sequence comparison is expected to contribute to a more appropriate construction of the initial tree and therefore at an overall improvement on multiple alignment of protein families. The initial tree is derived from the matrix of distances between separately aligned pairs of sequences, and based on experience,[5] it appears that unsuccessful multiple alignments are strongly dependent on errors made by initial alignments which produce an incorrect tree topology. Also, if the topology of the tree is correct, it has been remarked that mismatches among pairs of residues might propagate along different steps of the multiple alignment process and produce undesired results, especially for divergent sequences. PHYBAL good performance on pairs of divergent sequences is expected to allow a definite improvement on difficult pairwise alignment cases leading to the construction of a more reliable initial tree to be used in a classical multiple alignment approach.

Finally, the pairwise global alignment proposed by PHYBAL allows a pertinent comparison of structurally important regions predicted via hb and could be of great help for assessing the relationship between distantly related proteins. It has been shown that homologous proteins, even with < 20% identity, can be identified by estimating the overlap of rss in aligned sequences,[50] and similarly, an indicator of the overlap of hb between aligned sequences could be envisaged to discriminate between related and unrelated proteins.

ACKNOWLEDGMENTS

Part of this work has been done while AC and CD were visiting the Institut des Hautes Études Scientifiques.[51] AC is grateful to Jean-Michel Camadro for having suggested to look at the Hydrophobic Cluster Analysis approach. The authors thank Fréderic Dardel, Arthur Lesk, and Félix Rey for suggesting pairs of distantly related proteins to test which they included in the eight pairs dataset, Linda Dib and Dick Madden for carefully reading the manuscript.

REFERENCES

1. Darnell JE, Doolittle WF. Speculations on the early course of evolution. Proc Natl Acad Sci USA 1986;83:1271–1275.
2. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman D. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25:3389–3402.
5. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 1994;22:4673–4680.
6. Laget MP, Callebaut I, deLaunoit Y, Stehelin D, Mornon JP. Predicted common structural features of DNA-binding domains

from Ets, Myb and HMG transcription factors. Nucl Acids Res 1993;21:5987–5996.
7. Callebaut I, Renoir JM, Lebeau MC, Massol N, Burny A, Baulieu EE, Mornon JP. An immunophilin that binds M(r) 90,000 heat shock protein: main structural features of a mammalian p59 protein. Proc Natl Acad Sci USA 1992;89:6270–6274.
8. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins matrices for detecting distant relationships. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington, DC: National biomedical research foundation; 1978. 345–358.
9. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science 1992;256:1433–1445.
10. Henikoff S, Henikoff JG. Amino acids substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
11. Bashford D, Chothia C, Lesk AM. Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol 1987;196:199–216.
12. Bowie JU, Reidhaar-Olson JF, Wendell AL, Sauer RT. Deciphering the message in protein sequences:tolerance to amino acids substitution. Science 1990;247:1306–1310.
13. Overington JP, Donnelly D, Johnson MS, Sali A, Blundell TL. Environnement-specific amino acids substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1992;1:216–226.
14. Lesk A, Levitt M, Chothia C. Alignment of the amino acids sequences of distantly related proteins using variable gap penalties. Protein Eng 1986;1:77–78.
15. Teodorescu O, Galor T, Pillardy J, Elber R. Enriching the sequence substitution matrix by structural information. Proteins 2004;54:41–48.
16. Bowie JU, Luthy R, Einsenberg D. A method to identify protein sequences that fold into a known three dimensional structure. Science 1991;253:164–170.
17. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
18. Huang Y-M, Bystroff C. Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. Bioinformatics 2006;22:413–422.
19. Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem 1959;14:1–63.
20. Colon W, Elove GA, Wakem LP, Sherman F, Roder H. Side chain packing of the N- and C-terminal helices plays a critical role in the kinetics of cytochrome c folding. Biochemistry 1996;35:5538–5549.
21. Eliezer D, Yao J, Dyson HJ, Wright PE. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. Nat Struct Biol 1998;5:148–155.
22. Tsai J, Gerstein M, Levitt M. Simulating the minimum core for hydrophobic collapse in globular proteins. Protein Sci 1997;6: 2606–2616.
23. Dobson CM, Karplus M. The fundamentals of protein folding: bringing together theory and experiment. Curr Opin Struct Biol 1999;9:92–101.
24. Zdanowski K, Dadlez M. Stability of the residual structure in unfolded BPTI in different conditions of temperature and solvent composition measured by disulphide kinetics and double mutant cycle analysis. J Mol Biol 1999;287:433–445.
25. Hodsdon ME, Frieden C. Intestinal fatty acid binding protein: the folding mechanism as determined by NMR studies. Biochemistry, 2001;40:732–742.
26. Selvaraj S, Gromiha M. Role of hydrophobic clusters and long-range contact networks in the folding of α/β barrel proteins. Biophys J 2003;84:1919–1925.
27. Gaboriaud C, Bissery V, Benchetrit T, Mornon J-P. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett 1987;224:149–155.
28. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon J-P. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell mol life sci review 1997;53:7621–645.
29. Lemesle-Varloot L, Henrissat B, Gaboriaud C, Bissery V, Morgat A, Mornon J-P. Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences. Biochimie 1990;72:555–574.
30. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 1970;48:443–453.
31. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
32. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environnement-specific substitution tables and using structure dependant gap penalities. J Mol Biol 2001;310:243–257.
33. Stebbings LA, Mizuguchi K. HOMSTRAD: recent developments of the homologous protein structure alignment database. Nucl Acids Res 2004;32(Database issue):D203–D207.
34. Lackner F, Koppensteiner WA, Prlic MJ, Domingues FS. ProSup: a refined tool for protein alignment. Protein Eng 2000;13:745–752.
35. Prlic A, Domingues FS, Prlic MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.
36. Johnson MS, Overington JP. A structural basis for sequence comparison: an evaluation of scoring methodologies. J Mol Biol 1993;233:716–738.
37. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
38. Kawabata T. MATRAS: a program for protein 3D structure comparaison. Nucl Acids Res 2003;31:3367–3369.
39. Krissinel E, Henrick K. Protein structure coomparison in 3D based on secondary structure matching (SSM) followed by an alignment, scored by a new structural similarity function. 2003; Proceedings of the 5th International Conference on Molecular Structural Biology. Vienna, September 3–7, 2003.
40. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 2004;60:2256–2268.
41. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;297:1003–1013.
42. Thompson JD, Plewniak F, Poch O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 1999;15:87–88.
43. Hubbard SJ, Thornton JM. NACCESS VR. 1.1. Computer Program, Department of Biochemistry and Molecular Biology, University College London. 1993.
44. Kinjo AR, Nishikawa K. Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of mode of sequence conservation in proteins. Bioinformatics 2004;20:2504–2508.
45. Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 1996;264:823–838.
46. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
47. Poupon A, Jebai F, Labesse G, Gros F, Thibault J, Mornon JP, Krieger M. Structure modelling and site-directed mutagenesis of the rat aromatic L-amino acid pyridoxal 5′-phosphate-dependent decarboxylase: a functional study. Proteins 1999;37:191–203.
48. Carret C, Delbecq S, Labesse G, Carcy B, Precigout E, Moubri K, Schetters TP, Gorenflot A. Characterization and molecular cloning of an adenosine kinase from Babesia canis rossi. Eur J Biochem 1999;265:1015–1021.
49. Callebaut I, Prat K, Meurice E, Mornon J-P, Tomavo S. Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. BMC Genomics 2005;6:100.
50. Geourjon C, Combet C, Blanchet C, Deleage G. Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci 2001;10:788–797.
51. Deremble C. Détection de clusters hydrophobes à une dimension et alignements de séquences. Master thesis in "Modélisation Dynamique et Statistique des Systèmes Complexes," 2003, Université Pierre et Marie Curie, Paris.