Eur Biophys J (2007) 36:1059–1069
DOI 10.1007/s00249-007-0188-5

ORIGINAL PAPER

A reduced amino acid alphabet for understanding and designing protein adaptation to mutation

C. Etchebest · C. Benros · A. Bornot · A.-C. Camproux · A. G. de Brevern

Received: 13 February 2007 / Revised: 5 May 2007 / Accepted: 7 May 2007 / Published online: 13 June 2007

Abstract Protein sequence world is considerably larger than structure world. In consequence, numerous non-related sequences may adopt similar 3D folds and different kinds of amino acids may thus be found in similar 3D structures. By grouping together the 20 amino acids into a smaller number of representative residues with similar features, sequence world simplification may be achieved. This clustering hence defines a reduced amino acid alphabet (reduced AAA). Numerous works have shown that protein 3D structures are composed of a limited number of building blocks, defining a structural alphabet. We previously identified such an alphabet composed of 16 representative structural motifs (5-residues length) called Protein Blocks (PBs). This alphabet permits to translate the structure (3D) in sequence of PBs (1D). Based on these two concepts, reduced AAA and PBs, we analyzed the distributions of the different kinds of amino acids and their equivalences in the structural context. Different reduced sets were considered. Recurrent amino acid associations were found in all the local structures while other were specific of some local structures (PBs) (e.g Cysteine, Histidine, Threonine and Serine for the α-helix Ncap). Some similar associations are found in other reduced AAAs, e.g Ile with Val, or hydrophobic aromatic residues Trp with Phe and Tyr. We put into evidence interesting alternative associations. This highlights the dependence on the information considered (sequence or structure). This approach, equivalent to a substitution matrix, could be useful for designing protein sequence with different features (for instance adaptation to environment) while preserving mainly the 3D fold.

Keywords Amino acid classification · Structure-sequence relationship · Local protein structure · Secondary structure

Introduction

The large majority of proteins are composed of the classical 20 kinds of amino acids. This chemical diversity gives rise to a multitude of biological functions and a less extent to numerous structural folds. Thank to in vitro amino acid substitution performed in large-scale mutagenesis approaches, our knowledge and understanding of biological functions were considerably increased and improved (Buhot et al. 2004; Dubreuil et al. 2005).

However, such systematic studies require extremely tremendous work to assess the influence of each type of amino acid at each position on structural and functional properties of protein. Clearly, the appropriate selection of an amino acid type in a reliable set would be helpful to limit the number of experiments. Furthermore, experimental and theoretical studies have suggested that the full sequence complexity is not essential for the correct protein folding (Clarke 1995; Kuhlman and Baker 2004; Plaxco et al. 1998) and so, different works have been carried out to

Presented at the joint biannual meeting of the SFB-GEIMM-GRIP, Anglet France, 14–19 October, 2006.

C. Etchebest and C. Benros contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00249-007-0188-5) contains supplementary material, which is available to authorized users.

C. Etchebest · C. Benros · A. Bornot · A.-C. Camproux · A. G. de Brevern (✉)
Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S 726, Université Denis DIDEROT, Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France
e-mail: debrevern@ebgm.jussieu.fr

find the minimal amino acid alphabet (AAA). The simplest alphabet describes only two states: hydrophobic and polar. It has been used to design libraries of protein-like structures with some successes (Bradley et al. 2006, 2007; Hecht et al. 2004). It was also rather efficient to trap important features of the folding process when used with small sequences (Kamtekar et al. 1993; Regan and DeGrado 1988; Wei et al. 2003). For instance, based on lattice statistical mechanics, Dill's group developed theory to account for the folding of a heteropolymer molecule such as a protein to a globular and soluble state described with the simplest two-states alphabet (Dill 1985). General principles of protein structure, stability, and folding kinetics were so explored by computer simulations using simple exact lattice models involving few parameters, approximations, or implicit biases (Dokholyan 2005). They allowed complete explorations of conformational and sequence spaces (Dill et al. 1995; Sali et al. 1994). More importantly, it permitted to discuss on major driving forces of protein folding (Chan and Dill 1990).

Nevertheless, real proteins need more diversity than only two kinds of amino acid and these simple approaches have limitations (Yue et al. 1995), the diversity playing in particular a major role in the kinetic properties. Even if the number of experimental studies based on reduced AAA is limited, some are impressive. Riddle et al. (1997) designed a Src SH3 protein of 57 residues using an AAA composed of a limited set of 5 distinct amino acids (I, A, G, E and K). This combinatorial chemistry approach allowed the experimentalists to sample a wide range of possible mutations that could code for both ''foldability'' and function. Likewise, for the 213-residue *Escherichia coli* ororate phosphoribosyltransferase, 88% of the residues were changed with an AAA limited to nine amino acids (A, D, G, L, P, R, T, V and Y) (Akanuma et al. 2002). In this case, the reduction of the AAA was entirely supervised. Nevertheless, even with no influence on the protein topology and the conservation of the main biological functions, the reduction of the amino acid kinds influences the folding rate (Kuhlman and Baker 2004).

The common way to design a reduced AAA consists to cluster amino acids into groups according to specific features. These features may use sequence or structure information. A Substitution matrix such PAM or BLOSUM encountered in the field of sequence alignment is the most common usage of equivalence between amino acids. For instance, BLOSUM50 similarity matrix, based only on highly conserved regions in series of alignments without gaps (Henikoff and Henikoff 1992) was used by Murphy et al. (2000) to characterize a reduced AAA. Li et al. (2003) —with various clustering schemes—used BLOSUM62 and analyzed the consequences of the reduction of the number of amino acid kinds on sequence alignments. In the same way, Rogov and Nekrasov (2001), exploited the influence of the neighboring residues, while Smith and Smith (1990) simply analyzed aligned sequences.

Alternatively, some approaches used the structural information. Wang and Wang (1999) work relied on Miyazawa–Jernigan (MJ) matrix, i.e. an interaction potential matrix established from the analysis of a large set of 3D protein structures. In that case, interaction potential is defined between amino acids and is based on the observed frequency of contact of two amino acids in globular proteins (Miyazawa and Jernigan 1993). Depending on the clustering methods used, slightly different results were obtained (Cieplak et al. 2001; Esteve and Falceto 2004). In the same way, Solis and Rackovsky (2000) obtained an alphabet using information theory by reserving the *maximal information* in proteins described by backbone virtual bonds connecting consecutive C_α. Other methods have been tested, based for instance on the analysis of the amino acid distribution in secondary structures (Liu et al. 2003). The most recent developments have highlighted the necessity of conserving at least ten kinds of amino acids to ensure enough diversity (Fan and Wang 2003). All these different approaches often led to highly divergent concluding results indicating notably that the definition of a reduced AAA is highly dependent on the information used and the clustering method.

In most of these studies, no direct consideration of the influence of the local protein structures (Fitzkee et al. 2005) was introduced. However, it is now well accepted that protein structures can be seen as a combination of small local structures, or prototypes, yielding a more detailed description than classical secondary structures. A complete set of prototypes defines ''a structural alphabet'' that approximates accurately protein structures (Camproux et al. 1999, 2004; de Brevern et al. 2000; Karchin 2003; Sander et al. 2006; Unger et al. 1989). We proposed such a structural alphabet which is composed of 16 average protein fragments of 5 residues in length called Protein Blocks (PBs, see Supplementary data 1). We have limited the analysis to fragments of 5-residue length because it is sufficient to describe more than a short α-helix [four residues (Kumar and Bansal 1998)] and a minimal β structure [three residues (Colloc'h et al. 1993)].This alphabet was used both to describe 3D protein backbones but also to perform local structure prediction (de Brevern 2000, 2004, 2005, 2007; Etchebest et al. 2005). Moreover, PBs have proven their efficiency both in description and prediction of longer fragments (Benros 2005; Benros et al. 2006; de Brevern and Hazout 2003; de Brevern et al. 2002), loop conformations (Fourrier et al. 2004), to compare protein structures (Tyagi 2006; Tyagi et al. 2006a, b) and recently to detect magnesium-binding sites (Dudev and Lim 2007). The features of this alphabet were compared with those of eight other structural alphabets (Karchin et al. 2003). The

results have shown clearly that our PB alphabet is highly informative, with the best predictive ability of those tested (Karchin et al. 2003).

In this paper, we propose to use this structural alphabet to analyze equivalences between the different kinds of amino acids. By taking advantage of the description of every type of local protein structures, we analyze the relevance of the amino acid clusters obtained and discuss the different reduced sets of amino acids.

Materials and methods

Protein blocks

The 16 structural local prototypes are fragments of M (=5) residues long, corresponding to sequence windows of eight consecutive (ψ, ϕ) dihedral angles (de Brevern et al. 2000). The PB assignment is done using the *root mean square deviation on angular values*, i.e. an Euclidean distance on dihedral angles (ψ, ϕ) [see Supplementary data 1 and Fig. 1 of Etchebest et al. (2005)]. PBs m and d correspond to the prototypes describing central α-helix and central β-strand, respectively. PBs a through c primarily represent β-strand N-caps and e and f, C-caps. PBs g through j are specific to coils, k and l to α-helix N-caps, and n through p to α-helix C-caps (de Brevern 2005; de Brevern et al. 2000).

Data set

We used a set of protein structures derived from PDB-REPRDB composed of 1,407 protein chains and 293,507 residues (Noguchi and Akiyama 2003; Noguchi et al. 2001) taken from the PDB (Berman et al. 2000). The set contained proteins with no more than 30% pairwise sequence identity. We selected chains with a resolution better than 2.0 Å and a R-factor less than 0.2. Pairwise root mean square deviation (*rmsd*) values for all the chains are more than 10 Å.

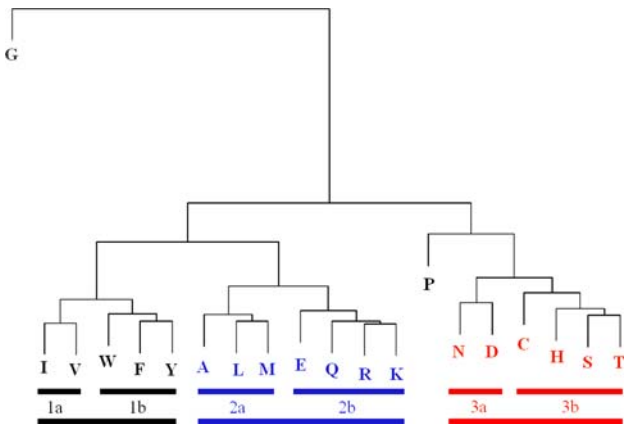

Fig. 1 *Amino acid clusters*. A hierarchical clustering using all the amino acid occurrence matrices of the 16 PBs was performed. From these amino acid distributions, three major clusters noted from 1 to 3 were defined. Each cluster was sub-divided into two sub-clusters with (*1a*): *I*, *V*; (1b): *W*, *F*, *Y*; (2a): *A*, *L*, *M*; (2b): *E*, *Q*, *R*, *K*; (3a): *N*, *D* and (3b): *C*, *H*, *S*, *T*. *P* and *G* were not considered in these clusters

Amino acid equivalence through a clustering analysis

We used the distribution of amino acids in PBs to create clusters of equivalent amino acids according to local structure. Once the databank was encoded in terms of PBs, sequence specificity was computed (see Supplementary data 2). Each PB was so associated with a set of enlarged sequence windows [$-w$; $+w$] of length l, (with $w = 7$ and $l = 15$). An amino acid occurrence matrix of dimension $20 \times l$ was computed for each PB. Then, each matrix was transformed into propensities matrix (de Brevern et al. 2000; Etchebest et al. 2005). Finally, all the matrices were compiled to create a matrix F of size $20 \times m$ with m, a vector of length $16l$ ($16 \times 15 = 240$). The distance D between two kinds of amino acids i and j was computed as follows:

$$D(\mathrm{aa}_i, \mathrm{aa}_j) = \sqrt{\sum_{k=1}^{m} (\mathrm{Faa}_i^k - \mathrm{Faa}_j^k)^2}$$

with Faa_i^k, the normalized frequency of amino acid i at position k. It corresponds to the frequency of amino acid i at position k divided by the frequency of amino acid i in the databank. Hierarchical clustering methods were then applied using R software (Ihaka and Gentleman 1996). Each resulting amino acid cluster represents amino acids that showed the same over- and under-representations upon all the PBs.

In a second step, we analyzed the behavior of the amino acid type in every PBs to ensure the relevance of the found clusters, i.e. if the residues in one cluster were always associated or not. In practice, the matrix F associated to the PB was used, in this case $m = l$. The use of Z-score matrices led to equivalent results.

Results

Global analysis

Figure 1 shows the final hierarchical clustering tree obtained. Glycine and Proline, two amino acids associated to specific local conformations of the protein backbone orientation, exhibited high specificities and could not be

associated with other amino acids. The remaining 18 amino acids were grouped into three clusters (noted 1, 2 and 3 in Fig. 1). For a more precise analysis, each one was split using supervised approach into two sub-clusters (noted a and b respectively). The clusters comprised from 5 to 7 amino acids. The first cluster, mainly hydrophobic, is composed of the sub-clusters: (1a) I, V (aliphatic amino acids) and (1b) F, Y, W (aromatic). The second cluster is more heterogeneous in terms of physico-chemical properties. It is composed of sub-clusters (2a) A, L, M (hydrophobic) and (2b) E, Q, R, K (polar/charged long amino acids). The last cluster includes polar amino acid with the sub-cluster (3a) N and D (polar/charged short amino acids) and the sub-cluster (3b) H, S, T and C (short and polar amino acids). We tested different analysis methods, e.g. Sammon Map (Sammon 1969) and Principal Component Analysis (Pearson 1901). Similar clusters were obtained which show that the results depend mainly on the specific amino acid distribution found in the different local structures. Only the branch length observed in Fig. 1 depends on the metric used. Interestingly, we can already notice that amino acids closely related by physico-chemical properties are not necessarily found co-associated. For instance, aliphatic residues Isoleucine (I) and Leucine (L) that differ only by one CH_2 group are associated to two different clusters (1a) and (2a) respectively. In the same way, Glutamate (E) and Aspartate (D), the two negatively charged polar amino acids—they differ also only by one CH_2 group—are found into two distinct groups (2b) and (3a) respectively. We have used an amino acid sequence window of length 15 similarly to our previous prediction works (de Brevern et al. 2000, 2004 2007; Etchebest et al. 2005). Nonetheless, it must be noted that the observed clustering is similar to the one found using a short amino acid sequence window of length 5 encompassing only the core of the PBs.

Local analysis

Table 1 gives the results obtained from the hierarchical clustering performed for each PB individually. In this table, the symbol (+) means that the set of amino acids regrouped in a given sub-cluster (or cluster) previously identified was preserved for the studied PB. Conversely, we observe changes within amino acid associations depending on the PBs. The considered associations of amino acids for each PB are given in brackets compared to the global analysis. This result enabled us to highlight the most stable clusters. With this approach, four different stability levels can be clearly observed:

(1) Three sub-clusters are highly stable, i.e., their amino acids are always associated whatever the PB. They correspond to the sub-clusters 1a (I and V), 1b (F, Y and W) and 3a (N and D). (2) The sub-cluster 2b (E, Q, R, K) is also highly stable with the only exception of PB *p*. In this PB, the Glutamic Acid (E) is not associated with the three other amino acids that of sub-cluster 2b. PB *p* is essentially found in protein local structures connecting α-helix to β-strand (see Supplementary data 1) and is characterized by a strong under-representation of Glutamic Acid in central position, i.e. the most informative one, and less important under-representation at the following positions. That could explain the absence of E in the sub-cluster 2b for this PB.

The two remaining sub-clusters are clearly less stable. (3) The sub-cluster 2a (A, L and M) is maintained for 11 of the 16 PBs. In four of the remaining PBs (PBs *c*, *e*, *h* and *i*), the Alanine is not grouped with Leucine and Methionine. For PB *j*, the Leucine (L) is not grouped with the two other amino acids. This last PB is the less frequent one, mainly associated with coil (see Supplementary data 1) and moreover is weakly structurally characterized, i.e. highest *rmsd* value of all the PBs (de Brevern 2005). In contrast, PBs related to helical structures (PBs *l* to *o*) present highly stable associations. (4) The sub-cluster 3b (H, S, T and C) is the least stable, i.e. the co-association of these amino acids for each PB is the least conserved. Histidine and Serine remains always associated in each PB and are found with the Threonine in 10 PBs. The only weak association is due to Cysteine. This latter is not found associated with the three others for five PBs (corresponding to α-helical structures and transitions from α-helix to β-strand, i.e. PB *n* to PB *b*) and only with the amino acids Histidine and Serine for five other PBs (PBs *c*, *d*, *e*, *h* and *i*).

Reducing the alphabet

All together, these results (see columns 1a + 1b, 2a + 2b and 3a + 3b of Table 1) show that the cluster 1 is the most stable because sub-clusters 1a and 1b are always preserved for all the PBs. It is not the case for clusters 2 and 3, because sub-clusters 2a and 3b are more versatile. Associations of the amino acids composing cluster 2 are thus only found for half of the PBs (PB *a*, *b*, *d*, *f*, *g*, *k* to *m*) and for cluster 3 for only five PBs (PB *f* and PBs *j* to *m*). In addition, another interesting point might be noticed: even if the association of H, C, S and T (sub-cluster 3b) is the most versatile one, interestingly Histidine and Serine are always found associated to the sub-cluster 3a. Clearly Cysteine and to a lesser extent Threonine behave differently depending on local protein structures examined. Hence, the AAA could be different depending on the PB considered.

Based on these observations, we can propose a reduced AAA based on sub-clusters 1a, 1b, 3a and eventually 2b. If we only consider amino acids always clustered together, we can reduce the AAA from 20 letters to 13 (see Fig. 2):

Table 1 Analysis of amino acid clusters for each PB

PB	Cluster 1			Cluster 2			Cluster 3			aa Differences	Major tendencies
	1a IV	1b FYW	1a + 1b	2a ALM	2b EQRK	2a + 2b	3a ND	3b HSTC	3a + 3b		
a	+	+	+	+	+	+	+	[*C*]/[H, S, T]	[*C*]/[H, S, T, N, D]	1	(1) + *C*/(2 + 3)
b	+	+	+	+	+	+	+	[*C*]/[H, S, T]	[*C*]/[H, S, T, N, D]	1	(1 + 2) + *C*/(3)
c	+	+	+	[*L*, *M*]/[A]	+	[*L*, *M*]/[A, E, Q, R, K]	+	[*C*, *T*]/[H, S]	[*C*, *T*]/[H, S, N, D]	4	(1) + [*L*, *M*, *C*, *T*]/(2 + 3)
d	+	+	+	+	+	+	+	[*C*, *T*]/[H, S]	[*C*, *T*]/[H, S, N, D]	2	(1) + [*C*, *T*]/(2 + 3)
e	+	+	+	[*L*, *M*]/[A]	+	[*L*, *M*]/[A, E, Q, R, K]	+	[*C*, *T*]/[H, S]	[*C*, *T*]/[H, S, N, D]	4	(1) + [*L*, *M*, *C*, *T*]/(2 + 3)
f	+	+	+	+	+	+	+	+	+	0	(1 + 2)/(3)
g	+	+	+	+	+	+	+	[*T*]/[*C*, H, S]	[*T*]/[C, H, S, N, D]	1	(1 + 2) + *T*/(3)
h	+	+	+	[*L*, *M*]/[A]	+	[*L*, *M*]/[A, E, Q, R, K]	+	[*C*, *T*]/[H, S]	[*C*, *T*]/[H, S, N, D]	4	(1) + [*L*, *M*, *C*, *T*]/(2 + 3)
i	+	+	+	[*L*, *M*]/[A]	+	[*L*, *M*]/[A, E, Q, R, K]	+	[*C*, *T*]/[H, S]	[*C*, *T*]/[H, S, N, D]	4	(1) + [*L*, *M*, *C*, *T*]/(2 + 3)
j	+	+	+	[*L*]/[A, M]	+	[*L*]/[A, M, E, Q, R, K]	+	+	+	1	(1) + [*L*]/(2 + 3)
k	+	+	+	+	+	+	+	+	+	0	(1 + 2)/(3)
l	+	+	+	+	+	+	+	+	+	0	(1 + 2)/(3)
m	+	+	+	+	+	+	+	+	+	0	(1 + 3)/(2)
n	+	+	+	+	+	[A, L, M]/[E, Q, R, K]	+	[*C*]/[H, S, T]	[*C*]/[H, S, T, N, D]	4	(1 + 2a) + [*C*]/(2b + 3)
o	+	+	+	+	+	[A, L, M]/[E, Q, R, K]	+	[*C*]/[H, S, T]	[*C*]/[H, S, T, N, D]	4	(1 + 2a) + [*C*]/(2b + 3)
p	+	+	+	+	[*Q*, *R*, *K*]/[E]	[*Q*, *R*, *K*]/[A, L, M, E]	+	[*C*]/[H, S, T]	[*C*]/[H, S, T, N, D]	4	(1 + 2) + [*C*]/(3) + [*Q*, *R*, *K*]

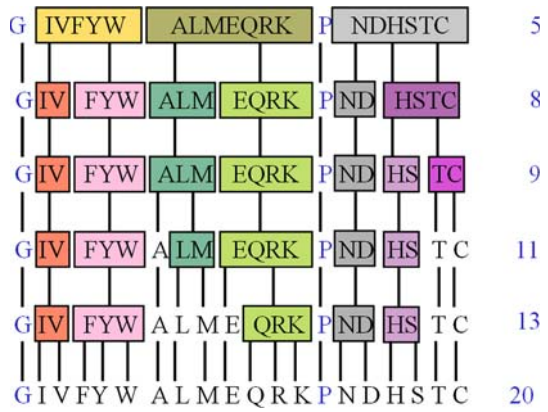

Fig. 2 *Different sets of amino acids.* The different sets of amino acids determined using our approach are shown in *color*

G, P, (I, V), (F, Y, W), A, L, M, E, (Q, R, K) (N, D), (H, S), T, C.

If some mismatches are allowed, i.e. ignoring the specific features in PBs, we can reduce the size of the alphabet furthermore. For instance, two mismatches i.e. sub-cluster 2a of PB *j* and sub-cluster 2b of PB *p*, lead to a final number of 11 amino acid types:

G, P, (I, V), (F, Y, W), A, (L, M), (E, Q, R, K) (N, D), (H, S), T, C.

This number can be reduced again to nine clusters of amino acids, with more mismatches:

G, P, (I, V), (F, Y, W), (A, L, M), (E, Q, R, K), (N, D), (H, S), (T, C).

Finally, in the smallest one, we can propose a five-letter alphabet, as often seen:

G, P, (I, V, F, Y, W), (A, L, M, E, Q, R, K), (N, D, H, S, T, C).

In this last case, amino acids with very different physico-chemical properties are grouped.

Discussion

The interest of defining a reduced AAA is based on the observation that an important number of amino acid substitutions have only limited effect on the final protein topology. The multiple-Alanine substitutions are one of the most explicit examples (Brown and Sauer 1999). However, even if the final topology and functions are conserved, the mutations often modify the folding rate and stability (Kuhlman and Baker 2004). In addition, the impact of the mutations is highly dependent on its location in the structure.

As it was already pointed out, depending on the objectives followed, the information used strongly influences the results (Esteve and Falceto 2005). Thus defining a reduced AAA based on the local structure may be more relevant than the sequence information alone. Indeed, numerous structure analysis and prediction methods have highlighted the usage of different amino acid kinds depending on the local protein structures (Aurora et al. 1997; Chothia et al. 1977; Fitzkee et al. 2005). We have presented here a new approach using directly the local description of protein structures through our structural alphabet. We have defined reduced AAAs and analyzed the potential equivalence between the amino acids. Our results highlight new interesting features that were not observed in previous studies.

As the number of reduced AAA is important, e.g., (Chan 1999; Dokholyan 2004; Li et al. 2003; Murphy et al. 2000; Xu and Miranker 2004), and their results diverging, we compared our results with two types of works, one based on the sequence information and the second one on structure information (see Table 2).

Sequence-based reduced alphabet

Murphy et al. (2000) used sequence alignments to create sets of reduced AAAs. Our results are partially in accordance. For instance, the residues composing cluster 1a (I, V) are also associated by their approach like the cluster 1b of (F, Y, W), and cluster 2b (E, Q, R, K). Nonetheless, some different associations are found which illustrate the influence of the information used, not related with the local structures. For instance, Alanine and Glycine are associated in the same group by Murphy et al., but not by our approach because these two amino acids are implicated in very different local protein structures. In the same way, Melo and Marti-Renom's work based on sequence alignments gave similar results to Murphy et al. (Melo and Marti-Renom 2006). The size of each group was very different corresponding respectively to 1, 1, 2, 7 and 9 kinds of amino acids. Clusters 1a and 1b remained associated while the nine-residues cluster corresponds to clusters 2b and 3a. Interestingly, H and C residues association that we found in *unstable* cluster 3b were isolated in their study. They also associated Alanine and Glycine in an independent group.

Table 2 Different reduced amino acid alphabets

Author's name	Year	Reduced alphabet
Riddle et al.	1997	IVFYWLMC AHT ED GP KNQRS
Wang and Wang	1999	
Akanuma et al.	1998	D P G A T V L R Y
Murphy et al.	2000	CILMV FYW AGPST DEHKNQR
Rogov and Nekrasov[a]	2001	M W C KRQE DNASTPGH VILFY
Esteve and Falceto	2004	STQNGPAHRED LIFVMYWCK
Melo and Marti-Renom	2005	C H AG FILMVWY DEKNPQRST

[a] For $m < 0.1$

Rogov and coworkers (Rogov and Nekrasov 2001) have analyzed the influence of the neighboring residues. Indeed, they showed that the mutual influence of amino acid residues is not limited to the nearest neighbours, but extends across significant distances in a polypeptide chain. The divergence with our results is less pronounced than for Murphy et al. For instance, the cluster 2b (E, Q, R, K) is found again like most of the sub-clusters. Nonetheless, in their analysis large clusters are created (see Fig. 3). The two major divergences concern the Methionine and Tryptophan that are not associated to another amino acid.

Structure-based reduced alphabet

The alphabet described by Baker's group and by Wang and Wang (Riddle et al. 1997; Wang and Wang 1999) is limited to five letters (I, A, G, E and K). Wang and Wang (1999) have used the MJ matrix (Miyazawa and Jernigan 1993) to reduce the amino acid kinds while Baker's results are based on SH3 domain experiments. At a first sight, generated amino acid clusters are different. Only cluster 1 is clearly found in both studies. In addition, with our approach, the two last amino acids (Glutamic Acid and Lysine) can be considered equivalent. Nonetheless, a deeper analysis shows stronger similarities. Firstly, the first cluster of Wang and Wang (I, V, F, Y, W, L, M, C) is found associated to some PBs (see Table 1), i.e. cluster (1) + [*L*, *M*, *C*]. In the same way, their second cluster including Serine and Threonine is similar to our cluster 3b. Their third cluster with Glutamine, Arginine and Lysine corresponds to our cluster 2b. In our approach, we have mainly a strong stable cluster 1 opposite to clusters 2 and 3. The Wang and Wang clusters also exhibit such separation.

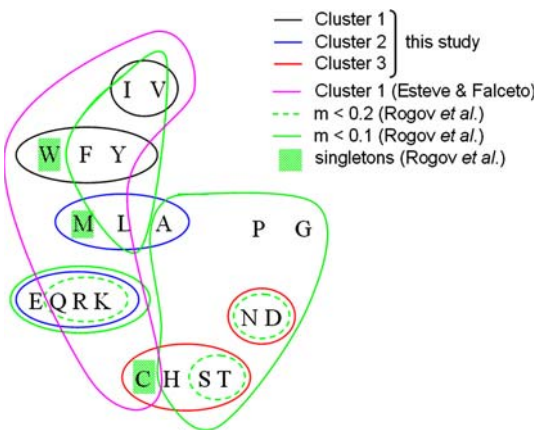

Fig. 3 *Representation of different amino acid associations.* The different sets of amino acids determined are shown (1) our work, in *black* (sub-clusters 1a and 1b), *blue* (sub-clusters 2a and 2b) and *red* (sub-clusters 3a and 3b), P and G are not associated to a cluster, (2) Esteve and Falceto work (Esteve and Falceto 2005), a clustering based on MJ matrix, in *pink* for the first cluster, all the other amino acids are associated to the second cluster and (3) Rogov and coworkers work (Rogov and Nekrasov 2001), a clustering based on aligned sequence, in *dashed green* a first level of clustering and in *plain green* a higher level defining three associations (Q, R and K), (S and T) and (N and D), in *green box* are highlighted the amino acids not associated to a cluster at any level (*m* is their measure of similarity)

In addition, it is important to notice that the MJ matrix often leads to a highly binary description. For instance, Esteve and Falceto using an unsupervised classification method based on subdominant ultrametric, defined two large amino acid clusters (Esteve and Falceto 2004). Comparison is thus difficult because we propose a larger number of smaller clusters. We can just note that no strong discordance is found; two large groups gathering our smaller groups, except for Alanine that is not associated to Methionine and Leucine (see Fig. 3). In our approach Alanine is not found associated to Leucine and Methionine for four PBs and so is considered as independent from (L, M) until the reduced amino acid set of nine clusters. This may be due to the peculiar behavior of Alanine which is often considered as a ''mimetic'' residue, i.e. it behaves as hydrophobic when surrounded with hydrophobic residues and reversely as a hydrophilic residue when embedded in a hydrophilic environment. This feature is used in the hydrophobic cluster analysis (Gaboriaud et al. 1987). This observation could also be related to the ''neutral'' character of Alanine with respect to mutation studies, such those performed in Alanine scanning experiments.

Conversely, our results are in accordance with the alphabet of nine amino acids used by Akanuma and coworkers (Akanuma et al. 2002) (G, P, V, Y, A, L, R, D and T). Only Alanine and Leucine separated by Akanuma and coworkers while we consider them as equivalent in our reduced AAA of nine kinds. It must be noted that Akanuma and coworkers have used both Alanine and Leucine in their initial amino acid subset because these two residues are the two most frequent amino acid of *E. coli* OPRTase, their target.

In most of the studies, physicochemical properties are not sufficient and are even misleading for associating two kinds of amino acids. For instance, Glutamate and Aspartate are negatively charged but associated to two different clusters. Our approach present the advantage to be based on local protein structures for defining clusters of equivalent amino acids. Hence, the reduced AAA defined in this way will tend to preserve the local structure and therefore the fold that maintains the function. Several authors have computed distributions associated to the different secondary structures (helix, strand, turn and coil), but only Liu et al. (2003) have exhaustively presented all the reduction steps. Unfortunately they did not analyze precisely the difference between the states, so no easy comparison with our results can be performed. Table 1 highlights the

importance of such an analysis and the complexity of amino acid association.

Protein design

Questions often arise about the usefulness of such description; we have presented in the introduction the application field of a reduced alphabet. Reduced AAA could help to design supervised mutations, protein design and prediction. It seems for instance highly suitable for *E. coli* OPRTase used by Akanuma et al. (2002). In the following, we propose examples of the interest of such approach.

Firstly, we analyzed the *N*-Carbamyl-D-amino acid amidohydrolase (*N*-carbamoylase) mutants 2S3 obtained by Oh and co-workers (Oh et al. 2002). This protein is employed in the industrial production of unnatural D-amino acid in conjunction with D-hydantoinase, but has low oxidative and thermostability. In this study, Oh and co-workers simultaneously tended to improve the oxidative and thermostability of *N*-carbamoylase from *Agrobacterium tumefaciens* NRRL B11291 by directed evolution using DNA shuffling. This work was in continuity of previous research on the same protein (Ikenaka et al. 1998a b; Nanba et al. 1998). They finally selected a mutant named 2S3, that greatly improved both oxidative and thermostability. It was purified and characterized. In this mutant, six amino acids were changed: Q23L, V40A, H58Y, G75S, M184L and T262A. To analyze the consequence of the sequence changes to the local structure, we have analyzed the structure of the highly homologous *N*-carbamoyl-D-amino-acid amidohydrolase from *A. radiobacter* [CRC14924, PDB ID: 1FO6 (Wang et al. 2001)]. The structure was first coded in terms of PBs and we then examined the observed mutations in terms of the clusters we previously defined. Table 3 summarizes the different data. The questions we address are (1) do the mutations belong to the same cluster and (2) is there any specificity in the local structure that would explain the change. A change from one sub-cluster to another very distinct cluster is clearly not equivalent to a slight change to a neighbouring cluster.

Table 3 Mutations of *N*-carbamoylase from *Agrobacterium tumefaciens* NRRL B11291

Position	Amino acid	Sub-cluster	Protein block	Mutated into	Sub-cluster
23	Q	2b	*m*	L	2a
40	V	1a	*a*	A	2a
58	H	3b	*c*	Y	1b
75	G	*out*	*k*	S	3b
184	M	2a	*m*	L	2a
262	T	3b	*f*	A	2a

For the *N*-carbamoylase from *A. tumefaciens*, only two mutations were selected in the same cluster (position 23 and 184), and four mutations correspond to a drastic change of clusters. For the position 40, it corresponds to a change from cluster 1a to cluster 2a (PB *a*), for the position 58 from 3b to 1b (PB *c*), for the position 75 from G to 3b (PB *k*), and for the position 262 from 3b to 2a (PB *f*). These changes of clusters correspond to a displacement to another cluster, but especially a cluster that is not associated to the original amino acid cluster for these local protein structures.

Positions 23 and 184 in 2S3 are Q23L and M184L are with PB *m*, a PB known to be extremely stable and associated to helical structures. Q23L resulted in the highest contribution to oxidative stability in the mutations found in 2S3, it is at the surface and it is only a change of sub-cluster from 2b to 2a. So, the change of amid group for a methyl is sufficient to improve the oxidative stability. The case of M184L is more complex as it is located near the enzyme core. Oxidation of Methionine residues is known to disrupt the protein structure (Kim et al. 2001). The suppression of the potential sulfoxide form of Methionine may give a stabilizing effect that can also be explained by increased hydrophobic interaction. The additional methyl group of Leucine could enhance hydrophobic interactions with F157 and V159 residues. We could conclude that changes of cluster are required for modifying the thermostability and oxidative features. In contrast, the preservation of stable structural elements is also necessary. It should be interesting to dispose of the whole set of data tested by the authors to better assessing the reliability and usefulness of our AAAs.

We performed another study based on the recent work done by Law and co-workers on the firefly luciferase of *Photinus pyralis*. This enzyme catalyzes a two-step reaction, using ATP-Mg^{2+}, firefly luciferin and molecular oxygen as substrates, leading to the efficient emission of yellow-green light (Law et al. 2006). They identified novel luciferase mutants which combine improved pH-tolerance and thermostability and retain specific activity of the wild-type enzyme (see Table 4). The effects of five amino acid replacements were additive (F14R, L35Q, V182K, I232K, F465R), and produced an enzyme with greatly improved pH-tolerance and stability. Combined mutant are superior to wild-type luciferase for many in vitro and in vivo applications.

To analyze the structure of luciferase, we have used the structure obtained by Franks et al. (1998) (PDB code: 1BA3). All mutations are found at the protein surface and lead to an amino acid of sub-cluster 2b (Q, R and K). It is mainly displacement from cluster 1a (V182K, I232K) and 1b (F14R, F465R) to 2b. In two cases, the change of

Table 4 Mutations of firefly luciferase of *Photinus pyralis*

Position	Amino acid	Sub-cluster	Protein block	Mutated into	Sub-cluster
14	F	1b	*b*	R	2b
35	L	2a	*m*	Q	2b
182	V	1a	*c*	K	2b
232	I	1a	*e*	K	2b
465	F	1b	*b*	R	2b

clusters is a displacement to a cluster not associated to the initial one (V182K for a PB *c* and I232K for PB *e*), while in the other cases, the new cluster is similar to the initial one (F14R and F465R both for PB *b*). Interestingly, residue 35, corresponding to a change from sub cluster 2a to 2b, is assigned to PB *m*.

Additional studies are required for assessing the hypothesis we propose about the requirement of selecting mutations in different clusters for improving the thermostability. However, such systematic analyses require a careful and exhaustive reading of the literature which is presently out of the scope of the present paper. In the same way, insofar as important structural elements are involved, mutations should be selected in the same cluster. These examples show that this kind of approach could greatly help experiments avoiding useless mutations.

In conclusion, this approach could be applied to design sequences highly compatible with a desired fold. Indeed, knowing a fixed series of PBs, i.e. protein fold, it is possible to find sequences able to adopt this given fold using statistical approaches such as the ones we have proposed (de Brevern et al. 2000; Etchebest et al. 2005). This reduced alphabet could be also useful in *threading* approach. Indeed, it could be used in a preliminary step of threading approach for detecting and selecting appropriate template. This approach could also be applied in alignment techniques as recently seen (Melo and Marti-Renom 2006; Wrabl and Grishin 2005). In the same way, it should be very useful when the sequence family is small or even if the sequence is orphan.

Acknowledgments This work was supported by French Institute for Health and Medical Care (INSERM) and University Paris 7-Denis Diderot. AB benefits from a grant of the Ministère de la Recherche.

References

Akanuma S, Kigawa T, Yokoyama S (2002) Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA 99:13549–13553

Aurora R, Creamer TP, Srinivasan R, Rose GD (1997) Local interactions in protein folding: lessons from the alpha-helix. J Biol Chem 272:1413–1416

Benros C (2005) Analyse et prédiction des structures tridimensionnelles locales des protéines, vol PhD. Paris 7-Denis Diderot, Paris, pp 212

Benros C, de Brevern AG, Etchebest C, Hazout S (2006) Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62:865–880

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Bradley LH, Thumfort PP, Hecht MH (2006) De novo proteins from binary-patterned combinatorial libraries. Methods Mol Biol 340:53–69

Bradley LH, Wei Y, Thumfort P, Wurth C, Hecht MH (2007) Protein design by binary patterning of polar and nonpolar amino acids. Methods Mol Biol 352:155–166

Brown BM, Sauer RT (1999) Tolerance of Arc repressor to multiple-alanine substitutions. Proc Natl Acad Sci USA 96:1983–1988

Buhot C, Chenal A, Sanson A, Pouvelle-Moratille S, Gelb MH, Menez A, Gillet D, Maillere B (2004) Alteration of the tertiary structure of the major bee venom allergen Api m 1 by multiple mutations is concomitant with low IgE reactivity. Protein Sci 13:2970–2978

Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Eng 12:1063–1073

Camproux AC, Gautier R, Tuffery P (2004) A hidden markov model derived structural alphabet for proteins. J Mol Biol 339:591–605

Chan HS (1999) Folding alphabets. Nat Struct Biol 6:994–996

Chan HS, Dill KA (1990) Origins of structure in globular proteins. Proc Natl Acad Sci USA 87:6388–6392

Chothia C, Levitt M, Richardson D (1977) Structure of proteins: packing of alpha-helices and pleated sheets. Proc Natl Acad Sci USA 74:4130–4134

Cieplak M, Holter NS, Maritan A, Banavar JR (2001) Amino acid classes and protein folding problem. J Chem Phys 114:1420–1423

Clarke ND (1995) Sequence 'minimization': exploring the sequence landscape with simplified sequences. Curr Opin Biotechnol 6:467–472

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

de Brevern AG (2005) New assessment of protein blocks. In Silico Biol 5:283–289

de Brevern AG, Hazout S (2003) 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19:345–353

de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41:271–287

de Brevern AG, Valadie H, Hazout S, Etchebest C (2002) Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 11:2871–2886

de Brevern AG, Benros C, Gautier R, Valadie H, Hazout S, Etchebest C (2004) Local backbone structure prediction of proteins. In Silico Biol 4:381–386

de Brevern AG, Etchebest C, Benros C, Hazout S (2007) ''Pinning strategy'': a novel approach for predicting the backbone structure in terms of protein blocks from sequence. J Biosci 32:51–70

Dill KA (1985) Theory for the folding and stability of globular proteins. Biochemistry 24:1501–1509

Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS (1995) Principles of protein folding—a perspective from simple exact models. Protein Sci 4:561–602

Dokholyan NV (2004) What is the protein design alphabet? Proteins 54:622–628

Dokholyan NV (2005) Studies of folding and misfolding using simplified models. Curr Opin Struct Biol 16:1–7

Dubreuil O, Bossus M, Graille M, Bilous M, Savatier A, Jolivet M, Menez A, Stura E, Ducancel F (2005) Fine tuning of the specificity of an anti-progesterone antibody by first and second sphere residue engineering. J Biol Chem 280:24880–24887

Dudev M, Lim C (2007) Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinformatics 8:106

Esteve JG, Falceto F (2004) A general clustering approach with application to the Miyazawa-Jernigan potentials for amino acids. Proteins 55:999–1004

Esteve JG, Falceto F (2005) Classification of amino acids induced by their associated matrices. Biophys Chem 115:177–180

Etchebest C, Benros C, Hazout S, de Brevern AG (2005) A structural alphabet for local protein structures: improved prediction methods. Proteins 59:810–827

Fan K, Wang W (2003) What is the minimum number of letters required to fold a protein? J Mol Biol 328:921–926

Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD (2005) Are proteins made from a limited parts list? Trends Biochem Sci 30:73–80

Fourrier L, Benros C, de Brevern AG (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 5:58

Franks NP, Jenkins A, Conti E, Lieb WR, Brick P (1998) Structural basis for the inhibition of firefly luciferase by a general anesthetic. Biophys J 75:2205–2211

Gaboriaud C, Bissery V, Benchetrit T, Mornon J-P (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett 224:149–155

Hecht MH, Das A, Go A, Bradley LH, Wei Y (2004) De novo proteins from designed combinatorial libraries. Protein Sci 13:1711–1723

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Ikenaka Y, Nanba H, Yajima K, Yamada Y, Takano M, Takahashi S (1998a) Increase in thermostability of N-carbamyl-D-amino acid amidohydrolase on amino acid substitutions. Biosci Biotechnol Biochem 62:1668–1671

Ikenaka Y, Nanba H, Yamada Y, Yajima K, Takano M, Takahashi S (1998b) Screening, characterization, and cloning of the gene for N-carbamyl-D-amino acid amidohydrolase from thermotolerant soil bacteria. Biosci Biotechnol Biochem 62:882–886

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680–1685

Karchin R (2003) Evaluating local structure alphabets for protein structure prediction, vol PhD. University of California, Santa Cruz, pp 301

Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51:504–514

Kim YH, Berry AH, Spencer DS, Stites WE (2001) Comparing the effect on protein stability of methionine oxidation versus mutagenesis: steps toward engineering oxidative resistance in proteins. Protein Eng 14:343–347

Kuhlman B, Baker D (2004) Exploring folding free energy landscapes using computational protein design. Curr Opin Struct Biol 14:89–95

Kumar S, Bansal M (1998) Geometrical and sequence characteristics of alpha-helices in globular proteins. Biophys J 75:1935–1944

Law GH, Gandelman OA, Tisi LC, Lowe CR, Murray JA (2006) Mutagenesis of solvent-exposed amino acids in Photinus pyralis luciferase improves thermostability and pH-tolerance. Biochem J 397:305–312

Li T, Fan K, Wang J, Wang W (2003) Reduction of protein sequence complexity by residue grouping. Protein Eng 16:323–330

Liu X, Zhang LM, Guan S, Zheng WM (2003) Distances and classification of amino acids for different protein secondary structures. Phys Rev E Stat Nonlin Soft Matter Phys 67:051927

Melo F, Marti-Renom MA (2006) Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins 63:986–995

Miyazawa S, Jernigan RL (1993) A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng 6:267–278

Murphy LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13:149–152

Nanba H, Ikenaka Y, Yamada Y, Yajima K, Takano M, Takahashi S (1998) Isolation of Agrobacterium sp. strain KNK712 that produces N-carbamyl-D-amino acid amidohydrolase, cloning of the gene for this enzyme, and properties of the enzyme. Biosci Biotechnol Biochem 62:875–881

Noguchi T, Akiyama Y (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 31:492–493

Noguchi T, Matsuda H, Akiyama Y (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). Nucleic Acids Res 29:219–220

Oh KH, Nam SH, Kim HS (2002) Improvement of oxidative and thermostability of N-carbamyl-D-amino acid amidohydrolase by directed evolution. Protein Eng 15:689–695

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2:559–572

Plaxco KW, Riddle DS, Grantcharova V, Baker D (1998) Simplified proteins: minimalist solutions to the 'protein folding problem'. Curr Opin Struct Biol 8:80–85

Regan L, DeGrado WF (1988) Characterization of a helical protein designed from first principles. Science 241:976–978

Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 4:805–809

Rogov SI, Nekrasov AN (2001) A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences. Protein Eng 14:459–463

Sali A, Shakhnovich E, Karplus M (1994) Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. J Mol Biol 235:1614–1636

Sammon JJW (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput 18:401–409

Sander O, Sommer I, Lengauer T (2006) Local protein structure prediction using discriminative models. BMC Bioinformatics 7:14

Smith RF, Smith TF (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. Proc Natl Acad Sci USA 87:118–122

Solis AD, Rackovsky S (2000) Optimized representations and maximal information in proteins. Proteins 38:149–164

Tyagi M (2006) New perspectives for protein structure analysis and mining using sequences of a structural alphabet, vol PhD. Université de la Réunion, Saint-Denis de la Réunion, pp 215

Tyagi M, Sharma P, Swamy C, Cadet F, Srinivasan N, de Brevern AG, Offmann B (2006a) Protein Block Expert (PBE): A web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res:W119–123

Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, Offmann B (2006b) A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 65:32–39

Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5:355–373

Wang J, Wang W (1999) A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 6:1033–1038

Wang WC, Hsu WH, Chien FT, Chen CY (2001) Crystal structure and site-directed mutagenesis studies of N-carbamoyl-D-amino-acid amidohydrolase from Agrobacterium radiobacter reveals a homotetramer and insight into a catalytic cleft. J Mol Biol 306:251–261

Wei Y, Kim S, Fela D, Baum J, Hecht MH (2003) Solution structure of a de novo protein from a designed combinatorial library. Proc Natl Acad Sci USA 100:13270–13273

Wrabl JO, Grishin NV (2005) Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. Proteins 61:523–534

Xu W, Miranker DP (2004) A metric model of amino acid substitution. Bioinformatics 20:1214–1221

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA (1995) A test of lattice protein folding algorithms. Proc Natl Acad Sci USA 92:325–329