Simplification of complexity in protein molecular systems by grouping amino acids: a view from physics

Jun Wang and Wei Wang

National Laboratory of Solid State Microstructure, Department of Physics, and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, China

ABSTRACT
Proteins are the main executors of complicated biological functions in nature. All the function-related structures and dynamics of proteins are encoded through the sequences with a 20-letter amino acid alphabet. Though the amino acid alphabets and the sequences are much more complex than that of the nucleic acids, the ubiquitous structural and functional homologs suggest there is much redundancy in the sequences and the alphabets. The discovery of the hidden simplicity of the amino acid alphabet could greatly help to disclose the regularities of protein systems and may stimulate some novel designs of protein, and has become one area in biological physics with continuous attentions. Physically, this kind of simplicity is a reflection and requirement of the folding landscape of proteins. In this review, we would like to describe the simplification problem of amino acid alphabet from physics aspect. The resultant extensions in bioinformatics and the relevant physical implications would also be discussed.
Classification is an important concept in various disciplines, such as biology, chemistry, physics, and even the various big-data applications nowadays. It provides a powerful way to elucidate the internal regularities inside the diverse data and phenomena [1]. For example, the classification of particles (known as the Eightfold Way or SU(3) flavor symmetry [2]) discloses the underlying regularities of the ‘particle zoo’, and predicted the fundamental degrees of freedom of matters [3]. The similar stories could be found frequently in scientific histories, such as the Dalton Atomic Theory, the Mendeleev Periodic Table of chemical elements, and so on. Now, the classification ideas and methods are also important in biological physics. Especially for the systems with only diverse and primitive knowledge, classification could help to extract the essence of the concerned systems. In this review, we would like to revisit the progresses on the classification of natural amino acids (namely the simplification of the amino acid alphabet), including the physical background, the methods, and the corresponding implications and applications in biophysics of proteins. These studies not only demonstrate the necessity of physics for such a problem, but also help to build better and concise physical models for the proteins. These researches could be an instructive example to study the biological objects with physical ideas and methods, and could be a window to understand various complicated biological behaviors.

1. Diversity and complexity of protein sequences

1.1. Proteins and amino acids

Proteins are a kind of polymers with essential biological functions [4]. As a connection between the biological information and functions, the protein is an irreplaceable component in central dogma of molecular biology [5]. Physically, the functions of proteins are rooted on their physico-chemical properties. This characteristic makes the protein an important entry to understand the connections between biology and physics. Presently, the proteins have become the common focus of the researches from various disciplines.

The basic philosophy of physics is the reductionism, that is to build the quantitative theory based on the properties of constituents. For protein systems, the amino acids are the basic building blocks of proteins. There are 20 kinds of standard amino acids which are common for various species, and could be encoded directly with genetic codes [4]. Through polymerization, these proteinogenic amino acids could be connected together to form chain-shaped polymers (i.e. proteins). In a certain protein, the order of amino acids along the chain is generally well specified due to natural selection or artificial designs, and is known as the primary structure of the proteins.

For many proteins, there is a correspondence between the amino acid sequences and their spatial structures under the standard condition, known as
the Anfinsen Principle [6]. That is, the proteins could autonomously fold the specific spatial structures with biological functions (called as native structure) guided by the physical interactions under the physiologic condition. This kind of relation between the composition and structure (function) supports the validity to understand the behaviors of proteins based on the properties of amino acids. Besides, there are practical conveniences to study proteins from the aspect of amino acids. Technically, the amino acid compositions and sequences of the proteins are easy to be obtained through chemical cleavages or DNA sequencing, while the determination of the structures or dynamics of proteins would be much more difficult [7] since small sizes (about tens of nanometers) and fast dynamics (span from nano-seconds to milli-seconds) are involved [8]. Considering the relationship between protein sequences and structures, to study structural and dynamic features of proteins from the aspect of amino acids would be an appealing way.

Along the direction, many important progresses on the sequence–structure relation of protein systems have been achieved during recent several decades [9–33], such as the identification and physical mechanisms of folding nuclei [10–13], the \textit{ab initio} methods in protein structural predictions [23–25], and the \textit{de novo} designs of functional proteins [26–29]. Surrounding the relationship between sequences and structures (functions) of proteins, these studies exposed the underlying quantitative rules of protein systems, build up the paradigm of protein physics and put it into practices, and thus greatly improve the knowledge about the protein systems [30–32].

1.2. Sequence–structure relationship and diversity of sequences

The sequence complexity is fundamental in protein systems [32,34]. For a typical size of protein domains, say 100 amino acids, there are about $20^{100} \approx 10^{130}$ possible sequences with the alphabet built with 20 kinds of amino acids. Fortunately, not all amino acid sequences are proteins. Based on the Anfinsen principle and the landscape theory, the native conformation is generally the most stable one thermodynamically, and there is a profound energy gap between the native state and the other conformations [6,31–33]. These rules act as the basic requirement for the protein sequences. Therefore, only a small part of these sequences could behave like proteins [32,35]. Yet, the possible number of proteins is still astronomical. For some sequences, their native structures are conservative with respect to the mutations of amino acids, while some behave like chameleons and may adopt various native conformations under different conditions [36–38]. The whole space built with these sequences is not uniform in their foldability and dynamic properties. What is the architecture of the sequence space? Can we use fewer kinds of amino acids to rebuild the protein ‘universe’? What would be the minimal size of the amino acid alphabet? The knowledge about the whole sequence space would bring people a global view of dynamics and evolution of protein systems [39].
Figure 1. The sketch for the effect of perturbation on energy landscape of protein system. Notes: After some kind of perturbations, the landscape of protein would change from the black one to the red one. For the stable system with large energetic bias (case (a)), the native conformation would be kept, while for the system with only marginal stability (case (b)), the native conformation would change from # to *.

Physically, the stability is often concurrently with the redundancy. For proteins, the structure and function could still be kept after a slight perturbation in environment or composition because of their stability (as shown in Figure 1). That is, for a protein system with a large energetic bias to its native conformation (# in Figure 1(a)), the native conformation (⋆ in Figure 1(a)) after a small perturbation (such as mutations) would still be the same as the original one (namely #). Differently, for a system with a marginal stability, the native conformation would be easily changed after some perturbations. This suggests that natural proteins with sufficient stability may tolerate many perturbations, and there are multiple sequences which share the same native structures. This is indeed observed in nature as well as in models. Based on the database of structural classification of proteins [40,41], the number of folds (corresponding to the structures) is much smaller than the number of domains (corresponding to sequences), which demonstrates the mapping from multiple sequences to one structure. With a model with two kinds of monomers, the sequence space could be characterized by the concept of voronoi cells in solid physics [42]. The center of these sequences is tightly related to the native structure. This outlines a basic architecture of sequence space. All these observations indicate that there is a large redundancy in sequence (composition) from the view of the structures.

Natural proteins are much more complicated. Some sequences with a similar structure could be rather different. In fact, this kind of phenomena are not rare [43,44]. For example, many proteins in a super-family have similar structural features and weak sequence similarities (sometimes may be smaller than 25%). In other words, a same structure could be encoded in totally different sequences. There might be multiple sequence clusters which could map to a structure. This
suggests that there are more complicated architectures in the sequence space, which deserves further careful studies.

1.3. How does diversity of sequences happen?

The complex relationship between sequences and structures originates from the compositional redundancy of natural proteins [45,46]. For the case with a small alphabet (for example, with only two kinds of monomers), the energy landscape of a related chain would generally be rugged, that is, there are many distinct structures whose energies are similar. It is difficult to design a foldable chain with such a small alphabet [47,48]. There would be tight relations between the native structure and the concerned sequences. Differently, when the alphabet is large enough, there are more degrees of freedom to stabilize a certain conformation without perturbations to other conformations. The resultant landscape could be rather smooth. Consequently, there would be more choices to design a funnel-shaped landscape targeting to a certain native conformation [49]. That is, for one structure, there might be multiple patterns of amino acids. Here, the different compositional redundancy might be the explanation of the difference between simplified models and the natural proteins. It seems that the natural proteins cannot be characterized with the model with just two kinds of monomers [50]. This would introduce a question: what size of the alphabet is minimal for a comprehensive description of protein systems? This question becomes one of the main concerns to decrypt the complexity of the protein sequences.

The diversity in the amino acid alphabet is also a result of natural evolution. During the evolution, the stability of the proteins may tolerate mutations of amino acids, which introduces the redundancy of the sequences even with a constant alphabet [39,51]. Besides, the addition of the new amino acids with innovative function groups could enhance the capability to build functional proteins and improve the fitness in evolution [52]. This kind of functional benefit would also favor the expansion of the proteinogenic alphabet. This demonstrates that the present alphabet is not random as the results of evolution [53]. Clearly, there are complex evolutionary factors in natural selections. The competition between these factors makes the regularities in the alphabet not straight-forward. To simplify the amino acid alphabet would demand serious considerations.

2. Simplification of amino acid alphabet

Toward deciphering sequence–structure relationship, many progresses have been achieved [9]. Experiments suggest that the diversity of the sequences could be simplified. For example, recent experiments demonstrated that the proteins (exemplified with the bovine pancreatic trypsin inhibitor) could be largely simplified with over one-third amino acids being alanine [54]. This kind of simplifications could help to find out the conservative sites in proteins. In this
sense, the reduction of the diversity of protein sequences is possible and helpful for protein studies. To reduce the diversity of the protein sequences, one of the powerful methods is to use simplified amino acid alphabets. This is coarse-graining modeling. By mapping the natural amino acids to simplified elements, the parameter space and the degrees of freedom of the systems could be greatly reduced, and the underlying physical rules could be highlighted. To realize such kinds of purposes, the simplification of amino acid alphabet is the fundamental step, and how to find out a rational simplification would be the central problem.

The simplification (classification) of amino acid alphabet is different from that for fundamental particles. On the one hand, there are a limited number of elements (presently 20 kinds of amino acids) in current alphabet. The existence of the detailed components is constrained by the ability of biological synthesis. Each amino acid in alphabet has to have its specific aminoacyl-tRNA synthetase on top of metabolic pathway for amino acids. The evolution of amino acid alphabet is tightly coupled with the variations of metabolic systems. Not every chemically available amino acid could be synthesized by the biological systems. This limits the possibility of novel elements, and large similarities may be observed between various amino acids. On the other hand, the current alphabet is tightly related to various kinds of diverse factors and processes [53], such as structural stability and biological functions. These factors and processes span a huge parameter space, and are often difficult to quantify. Additionally, the evolutions related to the various factors and processes are generally continuous. This makes the classes of elements rather fuzzy. All these considerations suggest that the simplification of amino acid alphabet is a complex problem.

2.1. Basic picture of grouping

Based on the contributions to a certain feature or function of proteins, the amino acids in the original alphabet $A_o$ could be classified into several groups. Typically, several amino acids with similar properties in the concerned process could be put into a group $G$. The whole set of these groups establishes a grouping $G = \{G\}$. In consequent applications, the amino acids in a group $G$ are represented with an element which possesses the common property. This common element may be artificially created. Sometimes, a certain amino acid in the group $G$ could be also chosen as the representative and be used to build simplified sequences. As a result, these artificial or selected representative elements would build up the simplified alphabet $A_s$. The number $N_G$ of groups $G$ in the grouping $G$ would give the size of the simplified alphabet $A_s$. Practically, it is assumed that there are no common elements shared by two groups in a certain grouping. This kind of separation generally suggests that the elements in various groups have sufficient differences in their concerned properties. This kind of classification could be operated iteratively, and correspondingly, a hierarchy of the groupings could be obtained.
Indeed, for a fixed original alphabet $A_o$, there are a huge number of groupings due to various kinds of combinations of elements [55]. There are total $5.1 \times 10^{13}$ combinations for all groupings, and $1.5 \times 10^{13}$ possible combinations for $N_G = 8$ case (the cases with the maximal number of combinations). Clearly, not all these groupings could satisfy the requirement of property separation. For a certain level of coarse-graining (such as with a certain $N_G$), there would be a few optimal groupings which could reproduce the features of systems with the original alphabet. These optimal groupings would be the targets in various studies on simplifications. Since $N_G$ is regularly an astronomical number for simplification problem of amino acids [55], an enumeration over all these combinations would be computationally consuming, and some sophisticated methods would be expected to find the optimal grouping. By quantifying the concerned properties, the amino acids are described in the property space. Sometimes, the optimal groupings could be constructed based on the geometrical relations in such a space. Here, the geometry ensures the optimality of the resultant grouping. Not all the possible groupings are considered, which introduces some efficiency gains. More generally, the optimal grouping could be derived by optimizing some certain target functions defined on the property space. Various kinds of optimization methods (such as genetic algorithm, Monte Carlo, and so on) could be used. Together with various methods, the definition of the geometry or the target functions would be key issues to find out the best simplifications.

### 2.2. Simplification based on design experiments

What is the best simplification? Clearly, the simplification which could successfully reproduce the structural and functional features of original proteins would be a good candidate. Therefore, to design the proteins based on a simplified alphabet would provide a way to check the validity of the simplified alphabet. This kind of idea has been realized in many experiments [56–63]. Facing with the helix bundles, de novo designs were realized based on hydrophobic/polar patterns along the protein chains [56,57]. The sequences of all the designed proteins are rather different, but share the same hydrophobic/polar pattern. This pattern is tightly related to the condition to be exposed or buried in the target helix-bundle structure. This pattern indicates, with a simplification with two kinds of elements, all these sequences are the same. The sequence–structure relationship is clearly exemplified for the four-helix-bundle fold, which demonstrates the power of the simplification of amino acid alphabet. Later, some bioinformatic and modeling studies also discovered HP patterns on the regular secondary structures (helix and strand) [64]. These directly prove the redundancy of the natural alphabet of amino acids as well as the possible simplifications for protein composition, and greatly benefit the design of protein designs [65].

Indeed, the simple hydrophobic/polar pattern is not the whole story of proteins. Further experiments found that, for a helix bundle, the cooperative thermal denaturation transition and hydrogen/deuterium exchange behavior
could be easily recovered with the reduced alphabets composed of three or seven kinds of amino acids [58,59]. More interestingly, a molecular design targeting on a beta-barrel-like structure (SH3 domain) was carried out with the combinatorial chemistry by Baker group [60]. The concerned structure is believed to be more complex than the helix bundle. It is found in their experiments that the alphabet with three kinds of amino acids fails to produce stable SH3 fold, and at least five kinds of amino acids are necessary. The minimal alphabet includes the amino acids \{I, A, G, E, K\}. This result gives out the minimal alphabet for the SH3 fold, and clearly supports the idea that the protein systems have more complexity than that indicated by simple HP models. To reproduce the catalytic activity, more types of amino acids are needed. Through sophisticated designs and selections, active enzymes could be constructed from 9-amino-acid alphabets [61–63]. It reflects that there are more demands with further function requirements. Indeed, the diversity of the resultant simplified alphabets suggests that to find out a universal simplified alphabet would be a difficult task since there are hundred kinds of folds in protein ‘universe’ and innumerable biological functions. Besides, how to map the original alphabet into this simplified one is still unknown. The minimal alphabet and the relation between the natural alphabet and the minimal one still need additional works.

2.3. Grouping based on physico-chemical features of individual amino acids

To analyze the physico-chemical features of individual amino acids would be a naive choice to derive the simplified alphabets. Though there are some additional factors (such as the connectivity and common backbone interactions) which affect the sequence–structure mapping, the properties of individual amino acids contribute essentially to the similarities and differences between proteins. To go from the features of individual amino acids is a good starting point to simplify amino acid alphabet.

The interaction is the basic essential one of amino acids. Among various interactions, the hydrophobic interactions contribute dominantly to the folding processes of proteins [18,66–68]. The hydrophobicity may act as the basic property for grouping. Quantitatively, based on the experimental free energy of transfer, the hydrophobicity scales of the hydrophobic and polar amino acids have different signs [69,70]. This explains molecule-level separation of these two kinds of amino acids in native conformation, and indicates the hydrophobic/polar division of alphabet. This is consistent with the experimentally suggested HP grouping.

Based on a certain scale, more groupings beyond the HP grouping could be derived. Considering the hydrophobic interaction, for example, all the amino acids could be placed along the axis describing the hydrophobic scales $H$ of amino acids. The distance between two kinds of amino acids could be determined as the difference of the scales of the concerned amino acids, $D_{ij} = |H_i - H_j|$. For the simplified alphabet, the distance between two groups $G_1$ and $G_2$ could be defined
Figure 2. The classification of amino acids based on their properties of side chains. Note: This is widely observed in literature [115].

based on $D_{ij}$ with $i(j) \in G_{1(2)}$ and. A regular definition could be $D(G_1, G_2) = \min_{i \in G_1, j \in G_2} (D_{ij})$. Accordingly, the amino acids could be easily divided into two groups based on the largest distance between the neighboring amino acids in this one-dimensional space. This gap clearly distinguishes two groups of amino acids with different properties. Practically, the distance $D(G_1, G_2)$ could be defined alternatively in other format to consider other aspects of the interactions and produce different groupings. This kind of division could be operated iteratively to the resultant groups. The hierarchy of the groups could be obtained.

However, it is difficult to create a single scale to describe various interactions in protein systems, since quantitative importance of various interactions is a delicate problem. To group the amino acids based on a specific interaction would generally overlook the effects of some others, and may produce artificial results. Actually, to go beyond the HP grouping, there are no clear pictures since other interactions are not well separated in their energy scales or populations. The corresponding groupings would interlace with each other (as shown in Figure 2). Even for different hydrophobic scales, the further groupings beyond the HP grouping would be rather different [71]. As a result, the practical groupings of amino acids would generally be based on the experiences or the features of concerned proteins. These kinds of choices sometime may introduce artificial biases to the researches.

Facing with the complexity of interaction, some proper projections would be necessary to grasp the main features of complex interactions. For example, the Miyazawa–Jernigen Matrix (MJ matrix) is a typical statistical potential [72–74], which contains the strengths of contact interactions between two kinds of amino acids in protein systems. Through the eigen analysis, this matrix could be represented as $M_{ij} \approx \lambda q_i q_j$, in which $M_{ij}$ is the element of MJ matrix, $\lambda$ and $\{q_i\}$ are the largest eigen value and the components of the
concerned eigen vector [75–78]. The eigen spectrum is shown in Figure 3(a). This gives out a representation of the pairwise interaction $M_{ij}$ with a one-body interaction scale $q_i$. This resembles the principal component analysis (PCA). Correspondingly, the grouping could be realized based on the resultant one-dimensional index (as shown in Figure 3(b)). For the MJ matrix, the HP grouping could be clearly identified based on this method. The simplification could be iteratively carried out based on this index [77,78]. Yet, when smaller differences between amino acids are involved, other eigen components should be considered since the approximation based on the principal components is not precise enough to characterize the differences of interactions. A high-dimension analysis is required based on the paradigm of PCA. This introduces much complexity in simplification. More generally, there are even no dominant components for some popular statistical potentials [79]. These limit the applications of the simplification based on the single-body properties of amino acids.

### 2.4. Simplification based on pairwise interaction between amino acids

Physically, pairwise potential is a better description of natural interaction, comparing to one-body scales. On the residue level, the strength of pairwise interactions could be determined statistically based on their occurrences in native structures, and are stored in the form of matrix [72–74,80–82]. The features of proteins have already been implicitly included. This kind of contact-based statistical interactions have been successfully applied in protein fold recognitions and structural predictions [73,83–88]. The simplification problem could be phrased as the simplification of the interaction matrix.

Different from informatic approaches [89], the simplified matrix should satisfy some physical requirements. Considering the foldability of proteins, it is expected that the simplified sequences of proteins based on the successful grouping could still fold to their native conformations. In the language of the energy landscape theory, this condition indicates that the energy of the simplified sequence in the native conformation of the original 20-letter sequence should still be the global minimum on the landscape of the simplified sequence [90]. Considering the excitation state with only one contact different from the native conformation, this condition could be reduced as the relation of contact energies, $\text{sgn}(\epsilon_{mn}^{\text{Orig}} - \epsilon_{kl}^{\text{Orig}}) = \text{sgn}(\epsilon_{AmBn}^{\text{Sim}} - \epsilon_{CkDl}^{\text{Sim}})$, in which $m(n,k,l)$ and $A_m(B_n,C_k,D_l)$ represents the type of the concerned amino acid and the related group. Due to the complexity of interaction, some combinations of $\{mn, kl\}$ may not satisfy this relation, and are defined as mismatches [90]. The content of mismatches for a certain grouping measures the quality of the simplification. With the mismatch as the target function, the optimal simplifications could be derived through the optimization over various groupings. Furthermore, some statistics could be carried out on the optimal groupings [79]. The frequency of certain groups could be used to measure their preferences in the rational grouping. A better hierarchy of rational groups could be produced based on the statistical information.
Figure 3. (a) Eigen spectrum of MJ matrix. There is a large eigen value which enables the PCA method. (b) The classification of amino acids based on the components of the dominant eigen vector of MJ matrix.
Note: Similar results are also obtained in the literature [78].
This procedure could be formulated as a problem of maximal likelihood estimation. For an interaction matrix $W_{ij}$, a hyper-matrix could be defined based on the difference between any two elements in $W$, $W_{mnkl} = W_{mn} - W_{kl}$. After the simplification, the element $W_{Sim}^{AmBnCkDl}$ of the simplified hyper matrix $W_{Sim}$ would correspond to a block $\{W_{mnkl}\}$ of elements in $W$. A successful simplification suggests the similarity between $W_{Sim}$ and $W$, that is, $W_{Sim}^{AmBnCkDl}$ and the elements in $\{W_{mnkl}\}$ should share some common features such as signs as defined above. Correspondingly, the similarity between $W_{Sim}$ and $W$ could be defined to measure the success of simplification, and the rational grouping then could be determined based on the maximal likelihood estimation. Similar to the distance matrix in the protein structural analysis, this description would be invariant to the transformation of the potential (such as the shift of the reference state) and ensures the validity of the final conclusions.

Based on the minimization of mismatches, not only the HP grouping is observed, but also some groupings with more complexity are obtained (as shown in Figure 4). For example, the optimal 5-letter alphabet is \{IAGEK\}, which is same as that observed in the molecular design experiment [60,91]. These groupings are further checked with simplified models [92,93]. With lattice models, a substantial number of 5-letter simplified sequences could fold to the native conformation of the original 20-letter sequences. Besides, the 5-letter substitutes could even reproduce the foldability and folding pathways of the original sequences. This is also observed in other simulations with high-resolution models [94]. More recently, the simplifications are further checked with more sophisticated computational models [95]. Focusing on several natural proteins and \textit{de novo} designed proteins, it is found that simplification using the 5-letter alphabet results in comparable quality of structure prediction to the full sequence of highly optimized proteins. The consistency implies the validity of the optimal groupings obtained through mismatch minimization.
2.5. **Simplification based on information in sequence**

Besides physical interactions, there are other data describing the relations between amino acids. The typical one is the scoring matrix (substitution matrix) used in sequence alignment [96,97], which measures the similarity between amino acids, and describes evolutionary relations of sequences. This offers another clue to simplify amino acid alphabets.

A simple way is to build the tree diagram using the similarity scores as the distances between amino acids [98]. The groupings could be built based on the resultant tree structure [99]. Besides, each line of the matrix could also be used as a vector $V^{(i)}$ describing a certain amino acid. In this sense, the classification problem could be phrased as a clustering problem in a 20-dimensional space. Based on certain definitions of distance in this space, the optimal groupings could be extracted [100,101]. These groupings grasp the main characteristics included in the scoring matrices, and correctly identify the features of the natural amino acid alphabet.

For protein systems, the sequence contains more information beyond that in amino acid alphabet. The good simplification should be able to identify the similarity between the original and simplified sequences. With the concept of global alignment, the similarity between two kinds of sequences could be expressed as the score $S = \sum_{E \in X} \sum_{R} \rho(R)B_{R,E}/N_X$, where $\rho(R)$ is the propensity of amino acid $R$, $B(R, E)$ is the similarity score between amino acids $R$ and $E$, and $N_X$ is the number of amino acids in the group $X$. Through the maximization of $S$, a series of optimal groupings could be obtained (as shown in Figure 5). It is interesting that the HP grouping determined with BLOSUM62 matrix is same as that from the interaction [102]. This consistency further confirms that the HP grouping is a fundamental feature of protein systems, and also demonstrates the correctness of these grouping methods which precisely characteristics of protein systems.

2.6. **Simplification based on features of protein systems**

Besides the analysis on the properties of amino acids, it is also possible to use the features of protein systems as the constraints for various simplifications. That is, the model proteins with simplified alphabets should be able to reproduce the features of the original proteins. Surely, there are a variety of features for protein systems. This idea of simplification would extend the view about the complexity inside protein systems.

The sequence–structure relation is a basic feature of proteins, and could be used as a typical requirement for simplification. This idea is implicitly included in other simplifications. Practically, the sequence–structure mapping could be described with information gains comparing to the random sequence [103], the ability of family recognition [104], as well as the mutual information based on contact population [105], and so on. Due to the complexity of the concerned data, some optimization methods are employed to search the optimal groupings which
Figure 5. The simplified groups which maximizes the information in simplified sequences. Note: Generated based on the data from [102].

can rebuild the properties of natural alphabet. One of the typical groupings is shown in Figure 6. It is found that the resultant groups embody various properties of amino acids, such as hydrophobicity/polarity, charge, size, aromaticity, and so on.

Besides the sequence–structure relation, there are still a large number of studies to simplify the amino acid alphabet based on various other features of proteins [106–114]. Though different features are concerned, their results of groupings are believed to be consistent with each other, and the physical interactions are considered to be the underlying determinant to produce these rational grouping [115]. This implies the importance of the physics in this bioinformatic problem.
2.7. Simplification based on genetic codes

In biological systems, the amino acid alphabet shares a common representation, that is, genetic codes. Each amino acid corresponds to one or more three-letter codes which correspond to the types of base groups of the sequential nucleic acids in the message RNA sequences. The codons and the amino acids correspond to the genotypes and phenotypes in the protein expression processes. Through the mutations on the level of nucleic acid, the types of expressed amino acids could be changed. Based on the symmetry analysis similar to those applied in particle physics, the simplified amino acid alphabet could also be derived[116–118].
In these studies, the degeneracy of the genetic codes for a certain amino acid could be the basic requirement. The chains of subgroups of various semisimple Lie algebra with 64 irreducible representations are systematically checked (as shown in Figure 7). It is found that only a certain chain of symmetry breaking could reproduce the degeneracy of the present codon table. Some other biological symmetries (such as the quartet or family box structure of the genetic code and its pyrimidine and purine exchange invariances) could also be included as the constraints. With these studies, not only the similarities between various amino acids could be identified, but also some evolutionary information of amino acid alphabet is suggested. It is interesting that there are no stages with only two kinds of amino acids, though the hydrophobic feature could be identified, which implies that the protein ‘universe’ is somehow created with sufficient complexity. Though simple, these results demonstrate the power of physics methods on the problem to group amino acids.

2.8. Minimal groupings of protein systems

With different sizes of alphabet, various optimal groupings correspond to different levels of coarse graining. These groupings generally include some information of natural proteins [115]. Similarities are observed for the groupings with various methods. [105, 115] Meanwhile, together with the simplifications, some features may not be correctly described after simplifications. For example, experiments and simulations suggest that the HP grouping could not describe all the features of natural proteins [92, 95] though it is a common feature of proteins. What are the minimal groupings for proteins? It is more fundamental in the simplification problem.

As indicated above, the quality of simplification groupings to model protein could be described with some target functions. It is found that there are some groupings which are preferred. For example, with the mismatch as the target function, there are some platforms (shoulders) on the curve of mismatch, around $N_G = 5$ and 8 (as shown in Figure 4). These platforms suggest that there are some local minima of mismatches, and could be viewed as certain stages of simplifications. For example, the 5-letter alphabet (IAGEK) has been successfully used in protein design, folding simulations and predictions. Related to the structural rebuilding, the minimal number of groups is around 5–6. While further considering the dynamic and functional behaviors, 8–10 groups of amino acids would be necessary [50, 79, 90, 92, 93, 102, 119–121]. Together with all these considerations, the minimal set to fulfill various kinds of biological functions might be the alphabet with 10 amino acids, which are verified with sequence design and homolog recognition [79, 93, 102, 105, 110, 122]. This measures the possible complexity of protein systems.
3. Implications of simplified amino acid alphabets

Through proper coarse graining, simplified amino acid alphabets could suppress the noises from the limited data and individual specificity, enhance the efficiency of bioinformatic analysis and molecular design, and bring a global view about the physical principle inside protein systems.

Firstly, the simplification could greatly enhance the capabilities to access large-scale sequence space. Practically, there are no sufficient data of proteins covering the whole sequence spaces. The statistical insufficiency of data is often an actual problem which may induce the biases due to data selections. The simplification could greatly reduce such kind of diversity in sequence analysis.
The concept of designability is a direct example [35,42,124]. The HP model enables the enumeration of the whole sequence space. The sequence–structure mappings are quantitatively illustrated, and the structure of the sequence space is then exemplified. This helps the further discussions about the evolutions of protein systems [39,51]. Recently, a more detailed study on the sequence–structure relation is also carried out based on binary models [125]. These studies with the simplest alphabet give us a better picture about the sequence–structure relation. In bioinformatics, the simplified amino acid alphabets are also widely used. The simplified alphabets are used, but not limited, in the prediction of protein–protein interaction and interface [126–128], nuclear receptors [129], DNA-binding proteins [130], defensin family and subfamily [131], heat shock protein family [132], residue flexibility [133,134], ability of protein crystallization [135] and various kinds of applications [136,137]. For example, the simplified alphabet (six groups) could be used to detect the sequence conservation related to the super-families of proteins [138,139]. It is well known that there are very weak similarities between 20-letter sequences in a super-family. This kind of coarse-graining helps to identify the sequential patterns originated from the structural conservations. Similarly, with a small-size alphabet (with about 10 letters), the ability to recognize the protein fold from simplified sequences could be enhanced [120]. These all demonstrate that the simplification could benefit the identification the hidden regularities in protein sequences. Indeed, this feature could be further observed for more complex protein systems (such as intrinsically disordered proteins). It is found that an alphabet with four kinds of amino acids could recognize the disorder with the accuracy similar to that with natural alphabet [140]. This also exemplifies the efficiency of simplified alphabet, and may help to disclose the physical connection between structural disorder and composition.

The simplified alphabets could also help to increase the efficiency in various kinds of researches, by reducing the complexity of the protein systems. This kind of benefit has been widely recognized in protein simulations. Various kinds of simplified models with reduced alphabets contribute importantly to the progresses of our knowledge about protein systems [141–145]. Yet, this kind of simplifications are often based on the intuitive groupings. The mapping between real proteins and the simplified models, especially the mapping between real sequences and the simplified ones, is still rare. The applications of simplified alphabets in computational designs and simulations would be a future direction. On the other hand, the enhancement of efficiency is widely observed in sequence-related bioinformatics studies. For example, a check for the accuracy and efficiency in the local homolog recognition with simplified alphabet is carried out [146]. It is found that the compressed alphabets could greatly improve the performance in local similarity discovery with a comparable coverage. Together with the measurement based on k-mer distance, the construction of the phylogenetic trees could even speed up for more than three orders of magnitude. Similar
results are also observed in other studies [147]. The efficiency gain could also be reached in experimental designs [148,149]. In a recent experiment, it is found that a design with 12 codons/12 amino acids could generate randomized libraries more ‘smarter’ with less screen effort comparing to that with 32 codons/20 amino acids [148]. They found that the libraries with simplified alphabet are of much higher quality, with the dramatically higher frequency of positive variants and the enhanced rate and enantioselectivity. These designs with simplified alphabets would enable an efficient search for the whole sequence space and could act as the valuable step to generate good candidates of functional proteins.

4. Summary

The simplification of alphabet of amino acids is an important way to reduce the complexity of protein sequences. The data describing the relations between amino acids are tightly related to their physical properties. Even when the biological information is considered, the methods and ideas from physics would also be essential to find out reasonable results. Here, we review some successful methods to simplify the amino acid alphabet. The connections between the physics and the grouping criterions are detailedly described. As a result, a series of grouping schemes are discovered and they successfully described many features of the simplified proteins. Due to the importance of the physical interactions in protein structure, dynamics and functions, there are some common features in various groupings. These results provide us tools to simplify and characterize the complexity in protein systems, and disclose the minimal simplicity inside the protein systems. The simplification processes demonstrate the importance of the underlying physics. There is generally a physical problem behind a bioinformatic problem. This could broaden our views to many bioinformatic researches and may stimulate many further studies.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Natural Science Foundation of China [grants number 11334004], [grants number 11174133], [grants number 81421091]; and National Basic Research Program of China [grant number 2013CB834100].

References

[1] H. Cohen and C. Lefebvre (eds.), Handbook of Categorization in Cognitive Science, Elsevier, Oxford, 2005.
[2] M. Gell-Mann, Synchrotron Laboratory Report CTSL-20 (CIT), 1961. Available at http://www.osti.gov/scitech/servlets/purl/4008239
[3] A. Zee, *The Search for Beauty in Modern Physics*, Princeton University Press, Princeton, NJ, 2007.

[4] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*, 6th ed., Garland Science, Taylor & Francis Group, New York, 2014.

[5] F. Crick, Nature 227 (1970) p.561.

[6] C.B. Anfinsen, Science 181 (1973) p.223.

[7] M.F. Juette, D.S. Terry, M.R. Wasserman, Z. Zhou, R.B. Altman, Q. Zheng and S.C. Blanchard, Curr. Opin. Chem. Biol. 20 (2014) p.103.

[8] H.S. Chan and K.A. Dill, Phys. Today 46 (1993) p.24.

[9] K.A. Dill and K.L. MacCallum, Science 338 (2012) p.1042.

[10] A. Fersht, *Structure and Mechanism in Protein Science*, Freeman, New York, 1999.

[11] A. Matouschek, J.T. Kellis, L. Serrano and A.R. Fersht, Nature 340 (1989) p.122.

[12] L.S. Itzhaki, D.E. Otzen and A.R. Fersht, J. Mol. Biol. 254 (1995) p.260.

[13] A.R. Fersht, Curr. Opin. Struct. Biol. 7 (1997) p.3.

[14] P.G. Wolynes, J.N. Onuchic and D. Thirumalai, Science 267 (1995) p.1619.

[15] J.N. Onuchic and P.G. Wolynes, Curr. Opin. Struct. Biol. 14 (2004) p.70.

[16] B.A. Shoemaker, J.J. Portman and P.G. Wolynes, Proc. Natl. Acad. Sci. USA 97 (2000) p.8868.

[17] P.E. Leopold, M. Montal and J.N. Onuchic, Proc. Natl. Acad. Sci. USA 89 (1992) p.8721.

[18] K.A. Dill, Biochemistry 29 (1990) p.7133.

[19] K.A. Dill and H.S. Chan, Nat. Struct. Biol. 4 (1997) p.10.

[20] K.A. Dill, S.B. Ozkan, M.S. Shell and T.R. Weikl, Annu. Rev. Biophys. 37 (2008) p.289.

[21] K.W. Plaxco, K.T. Simons and D. Baker, J. Mol. Biol. 277 (1998) p.985.

[22] D. Baker, Nature 405 (2000) p.39.

[23] D. Baker and A. Sali, Science 294 (2001) p.93.

[24] K.T. Simons, R. Bonneau, I. Ruczinski and D. Baker, Proteins: Struct. Funct. Bioinf. 37 (1999) p.171.

[25] A.N. Adhikari, K.F. Freed and T.R. Sosnick, Proc. Natl. Acad. Sci. USA 109 (2012) p.17442.

[26] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard and D. Baker, Science 302 (2003) p.1364.

[27] L. Jiang, E.A. Althoff, F.R. Clemente, L. Doyle, D. Rthlisberger, A. Zanghellini, J.L. Gallaher, J.L. Betker, F. Tanaka, C.F. Barbas 3rd, D. Hilvert, K.N. Houk, B.L. Stoddard and D. Baker, Science 319 (2008) p.1387.

[28] P. Bradley, K.M.S. Misura and D. Baker, Science 309 (2005) p.1868.

[29] J.B. Siegel, A. Zanghellini, H.M. Lovick, G. Kiss, A.R. Lambert, J.L. St. Clair, J.L. Gallaher, D. Hilvert, M.H. Gelb, B.L. Stoddard, K.N. Houk, F.E. Michael and D. Baker, Science 329 (2010) p.309.

[30] R. Nussinov and P.G. Wolynes, Phys. Chem. Chem. Phys. 16 (2014) p.6321.

[31] H. Frauenfelder, S.G. Sligar and P.G. Wolynes, Science 254 (1991) p.1598.

[32] J.N. Onuchic, Z. Luthey-Schulten and P.G. Wolynes, Annu. Rev. Phys. Chem. 48 (1997) p.545.

[33] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. 84 (1987) p.7524.

[34] A. Currin, N. Swainston, P.J. Day and D.B. Kell, Chem. Soc. Rev. 44 (2015) p.1172.

[35] H. Li, R. Helling, C. Tang and N. Wingreen, Science 273 (1996) p.666.

[36] W.Q. Zou and P. Gambetti, Cell. Mol. Life Sci. 64 (2007) p.3266.

[37] V.N. Uversky, J. Biomol. Struct. Dyn. 21 (2003) p.211.

[38] W.J. Anderson, L.O. Van Dorn, W.M. Ingram and M.H. Cordes, Protein Eng. Des. Sel. 24 (2011) p.765.
[39] T. Sikosek and H.S. Chan, J. R. Soc. Interface 11 (2014) p.20140419.
[40] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, J. Mol. Biol. 247 (1995) p.536.
[41] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha and A. Murzin, Nucleic Acids Res. 42 (2014) p.D310.
[42] H. Li, C. Tang and N.S. Wingreen, Proc. Natl. Acad. Sci. USA 95 (1998) p.4987.
[43] B. Rost, Folding & Design 2 (1997) p.S19.
[44] R.F. Doolittle, TIBS 19 (1994) p.15.
[45] P.G. Wolynes, Nat. Struct. Biol. 4 (1997) p.871.
[46] M.S. Shell, P.G. Debenedetti and A.Z. Panagiotopoulos, Proteins 62 (2006) p.232.
[47] T.M. Fink and R.C. Ball, Phys. Rev. Lett. 87 (2001) p.198103.
[48] R.C. Ball and T.M. Fink, Phys. Rev. E 66 (2002) p.031902.
[49] E. Munoz and M.W. Deem, Protein Eng. Des. Sel. 21 (2008) p.311.
[50] F. Melo and M.A. Marti-Renom, Proteins 63 (2006) p.986.
[51] R. Wroe, E. Bornberg-Bauer and H.S. Chan, Biophys. J. 88 (2005) p.118.
[52] M.M. Muller, J.R. Allison, N. Hongdilokkul, L. Gaillon, P. Kast, W.F. van Gunsteren, P. Marliere and D. Hilvert, PLoS Genet. 9 (2013) p.e1003187.
[53] G.K. Philip and S.J. Freeland, Astrobiology 11 (2011) p.235.
[54] M.M. Islam, S. Sohya, K. Noguchi, M. Yohda and Y. Kuroda, Proc. Natl. Acad. Sci. USA 105 (2008) p.15334.
[55] N. Cannata, S. Toppo, C. Romualdi and G. Valle, Bioinformatics 18 (2002) p.1102.
[56] S. Kamtekar, J. Schiffer, H. Xiong, J. Babik and M. Hecht, Science 262 (1993) p.1680.
[57] Y. Wei, S. Kim, D. Fela, J. Baum and M.H. Hecht, Proc. Natl. Acad. Sci. USA 100 (2003) p.13270.
[58] A.R. Davidson, K.J. Lumb and R.T. Sauer, Nat. Struct. Biol. 2 (1995) p.856.
[59] C.E. Schafmeister, S.L. LaPorte, L.J.W. Miercke and R.M. Stroud, Nat. Struct. Biol. 4 (1997) p.1039.
[60] D.S. Riddle, J.V. Santiago, S.T. Bray-Hall, N. Doshi, V.P. Grantcharova, Q. Yi and D. Baker, Nat. Struct. Biol. 4 (1997) p.805.
[61] J.A. Silverman, R. Balakrishnan and P.B. Harbury, Proc. Natl. Acad. Sci. USA 98 (2001) p.3092.
[62] S. Akanuma, T. Kigawa and S. Yokoyama, Proc. Natl. Acad. Sci. USA 99 (2002) p.13549.
[63] K.U. Walter, K. Vamvaca and D. Hilvert, J. Biol. Chem. 280 (2005) p.37742.
[64] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh and H.C. Lee, Phys. Rev. Lett. 84 (2000) p.386.
[65] L.H. Bradley, Methods Mol. Biol. 1216 (2014) p.117.
[66] C. Tanford, Science 200 (1978) p.1012.
[67] C. Tanford, The Hydrophobic Effect: Formation of Micelles and Biological Membranes, Wiley, New York, 1973.
[68] W. Kauzmann, Adv. Protein Chem. 14 (1959) p.1.
[69] W.C. Wimley, T.P. Creamer and S.H. White, Biochemistry 35 (1996) p.5109.
[70] S.H. White and W.C. Wimley, Annu. Rev. Biophys. Biomol. Struct. 28 (1999) p.319.
[71] J. Janin, Nature 277 (1979) pp.491–492; R. Wolfenden, L. Andersson, P.M. Cullis, and C.C.B. Southgate, Biochemistry 20 (1981) pp. 849–855; J. Kyte, and R.F. Doolittle, J. Mol. Biol. 157 (1982) pp. 105–132; G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee, and M.H. Zehfus, Science 229 (1987) pp. 834–838.
[72] S. Miyazawa and R.L. Jernigan, Macromolecules 18 (1985) p.534.
[73] S. Miyazawa and R.L. Jernigan, J. Mol. Biol. 256 (1996) p.623.
[74] S. Miyazawa and R.L. Jernigan, Proteins 34 (1999) p.49.
[75] H. Li, C. Tang and N.S. Wingreen, Phys. Rev. Lett. 79 (1997) p.765.
[76] Z.H. Wang and H.C. Lee, Phys. Rev. Lett. 84 (2000) p.574.
[77] M. Cieplak, N.S. Holter, A. Maritan and J.R. Banavar, J. Chem. Phys. 114 (2000) p.1420.
[78] H.S. Chan, Nat. Struct. Biol. 6 (1999) p.994.
[79] J. Wang and W. Wang, Phys. Rev. E 65 (2002) p.041911.
[80] P.D. Thomas and K.A. Dill, J. Mol. Biol. 257 (1996) p.457.
[81] M.R. Betancourt and D. Thirumalai, Protein Sci. 8 (1999) p.361.
[82] T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frelsen, C. Andreutta, W. Boomsma, S. Bottaro and J. Feringhing-Borg, PLoS One 5 (2010) p.e13714.
[83] D. Tobi and R. Elber, Proteins 41 (2000) p.40.
[84] M.Y. Shen and A. Sali, Protein Sci. 15 (2006) p.2507.
[85] P. Narang, K. Bhushan, S. Bose and B. Jayaram, J. Biomol. Struct. Dyn. 23 (2006) p.385.
[86] M.J. Sippl, Proteins 17 (1993) p.355.
[87] S.H. Bryant and C.E. Lawrence, Proteins 16 (1993) p.92.
[88] S. Kmiecik and A. Kolinski, Proc. Natl. Acad. Sci. USA 104 (2007) p.12330.
[89] S.J. Shepherd, C.B. Beggs and S. Jones, Euro. Biophys. J. Biophys. Lett. 37 (2007) p.105.
[90] J. Wang and W. Wang, Nat. Struct. Biol. 6 (1999) p.1033.
[91] J.G. Saven, Chem. Rev. 101 (2001) p.3113.
[92] J. Wang and W. Wang, Phys. Rev. E 61 (2000) p.6981.
[93] K. Fan and W. Wang, J. Mol. Biol. 328 (2003) p.921.
[94] E. Guarnera, R. Pellarin and A. Callisich, Biophys. J. 97 (2009) p.1737.
[95] H.H. Truong, B.L. Kim, N.P. Schafer and P.G. Wolynes, J. Chem. Phys. 139 (2013) p.121908.
[96] M.O. Dayhoff, R. Schwartz and B.C. Orcutt, A model of evolutionary change in proteins, in Atlas of Protein Sequence and Structure Supplement 3 ed., Vol. 5, M.O. Dayhoff, ed., National Biomedical Research Foundation, Silver Spring, MD, 1978, p.345.
[97] S. Henikoff and J.G. Henikoff, Proc. Natl. Acad. Sci. USA 89 (1992) p.10915.
[98] J. Felsenstein, Cladistics 5 (1989) p.164.
[99] A. Prlic, F.S. Domingues and M.J. Sippl, Protein Eng. 13 (2000) p.545.
[100] L.R. Murphy, A. Wallqvist and R.M. Levy, Protein Eng. 13 (2000) p.149.
[101] X. Liu, D. Liu, J. Qi and W.M. Zheng, Phys. Rev. E 66 (2002) p.021906.
[102] T.P. Li, K. Fan, J. Wang and W. Wang, Protein Eng. 16 (2003) p.323.
[103] A.D. Solis and S. Rackovsky, Proteins 38 (2000) p.149.
[104] J. Lenckowski and K. Walczak, Simplifying amino acid alphabets using a genetic algorithm and sequence alignment, in Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, E. Marchiori, J.H. Moore and J.C. Rajapakse, eds., EvoBIO 2007, LNCS 4447, Springer-Verlag, Berlin Heidelberg, 2007, p.122.
[105] A.D. Solis, Proteins 83 (2015) p.2198.
[106] G.M. Crippen, Biochemistry 30 (1991) p.4232.
[107] V.N. Maiorov and G.M. Crippen, J. Mol. Biol. 227 (1992) p.876.
[108] C. Etchebest, C. Benros, A. Bornot, A.C. Camproux and A.G. de Brevern, Eur. Biophys. J. 36 (2007) p.1059.
[109] I.B. Kuznetsov and S. Rackovsky, Proteins 49 (2002) p.266.
[110] J. Bacardit, M. Stout, J.D. Hirst, A. Valencia, R.E. Smith and N. Krasnogor, BMC Bioinf. 10 (2009) p.6.
[111] S. Pape, F. Hoffgaard, M. Duur and K. Hamacher, J. Comput. Chem. 34 (2013) p.1.
[112] M.N. Davies, A. Secker, A.A. Freitas, E. Clark, J. Timmis and D.R. Flower, Bioinfomatics 24 (2008) p.1980.
[113] E. Susko and A.J. Roger, Mol. Biol. Evol. 24 (2007) p.2139.
[114] J.O. Wrabl and N.V. Grishin, Proteins 61 (2005) p.523.
[115] J.D. Stephenson and S.J. Freeland, J. Mol. Evol. 77 (2013) p.159.
[116] J.D. Bashford, I. Tsokhantjis and P.D. Jarvis, Proc. Natl. Acad. Sci. USA 95 (1998) p.987.
[117] J.E.M. Hornos and Y.M.M. Hornos, Phys. Rev. Lett. 71 (1993) p.4401.
[118] A. Kawahara-Kobayashi, A. Masuda, Y. Araiso, Y. Sakai, A. Kohda, M. Uchiyama, S. Asami, T. Matsuda, R. Ishitani, N. Dohmae, S. Yokoyama, T. Kigawa, O. Nureki and D. Kiga, Nucleic Acids Res. 40 (2012) p.10576.
[119] A. Fernandez, Z. Naturforsch C. 59 (2004) p.151.
[120] E.L. Peterson, J. Kondev, J.A. Theriot and R. Phillips, Bioinformatics 25 (2009) p.1356.
[121] L. Chang, J. Wang and W. Wang, Phys. Rev. E 82 (2010) p.051930.
[122] L.M. Longo and M. Blaber, Arc. Biochem. Biophys. 526 (2012) p.16.
[123] A.J. Sodt and T. Head-Gordon, Biophys. J. 99 (2010) p.227.
[124] E.G. Emberly, N.S. Wingreen and C. Tang, Proc. Natl. Acad. Sci. USA 99 (2002) p.11163.
[125] E. Ferrada, PLoS Comput. Biol. 10 (2014) p.e1003946.
[126] J.W. Shen, J. Zhang, X.M. Luo, W.L. Zhu, K.Q. Yu, K.X. Chen, Y.X. Li and H.L. Jiang, Proc. Natl. Acad. Sci. USA 104 (2007) p.4337.
[127] Q. Huang, Z. You, X. Zhang and Y. Zhou, Int. J. Mol. Sci. 16 (2015) p.10855.
[128] G. Launay, R. Mendez, S. Wodak and T. Simonson, BMC Bioinf. 8 (2007) p.270.
[129] H.C. Wang and X.H. Hu, BMC Bioinf. 16 (2015) p.402.
[130] L. Nanni and A. Lumini, Amino Acids 36 (2009) p.167.
[131] Y.C. Zu, Y. Lv, Z.Y. Wei, L. Yang, G.P. Li and G.L. Fan, PLoS One 10 (2015) p.e0145541.
[132] P.M. Feng, W. Chen, H. Lin and K.C. Chou, Anal. Biochem. 442 (2013) p.118.
[133] H. Zhang and L. Kurgan, Amino Acids 46 (2014) p.2665.
[134] P. Craveur, A.P. Joseph, J. Esque, T.J. Narwani, N. Shinada, M. Goguet, S. Leonard, P. Poulain, O. Bertrand, G. Faure, J. Rebehmed, A. Ghozlane, L.S. Swapna, R.M. Bhaskara, J. Barnoud, S. Teletchea, V. Jallu, J. Cerny, B. Schneider, C. Etchebest, N. Srinivasan, J.C. Gelly and A.G. de Brevern, Front. Mol. Biosci. 2 (2015) p.20.
[135] P. Smialowski, T. Schmidt, J. Cox, A. Kirschner and D. Frishman, Proteins 62 (2006) p.343.
[136] M. Randic, J. Zupan, A.T. Balaban, D. Viktor-Topic and D. Plavsic, Chem. Rev. 111 (2011) p.790.
[137] M. Fernandez, J. Caballero, L. Fernandez and A. Sarai, Curr. Bioinf. 5 (2010) p.241.
[138] L. Mirny and E. Shakhnovich, J. Mol. Biol. 308 (2001) p.123.
[139] L.A. Mirny and E. I. Shakhnovich, J. Mol. Biol. 291 (1999) p.177.
[140] E.A. Weathers, M.E. Paulaitis, T.B. Woolf and J.H. Hoh, FEBS Lett. 576 (2004) p.348.
[141] A.I. Jewett, A. Baumketner and J.E. Shea, Proc. Natl. Acad. Sci. USA 101 (2004) p.13192.
[142] J.D. Honeycutt and D. Thirumalai, Biopolymers 32 (1992) p.695.
[143] J.E. Shea, J.N. Onuchic and C.L. Brooks III, J. Chem. Phys. 113 (2000) p.7663.
[144] C. Wu and J.E. Shea, Curr. Opin. Struct. Biol. 21 (2011) p.209.
[145] M.S. Li, M.T. Co, G. Reddy, C.K. Hu, J.E. Straub and D. Thirumalai, Phys. Rev. Lett. 105 (2010) p.218101.
[146] R.C. Edgar, Nucleic Acids Res. 32 (2004) p.380.
[147] H. Ogul and E.U. Mumcuoglu, Biosystems 87 (2007) p.75.
[148] M.T. Reetz, D. Kahakeaw and R. Lohmer, ChemBioChem 9 (2008) p.1797.
[149] Z. Sun, R. Lonsdale, X.D. Kong, J.H. Xu, J. Zhou and M.T. Reetz, Angew. Chem. Int. Ed. Engl. 54 (2015) p.12410.