SCIENTIFIC REPORTS
nature research

OPEN

Protein design under competing conditions for the availability of amino acids

Francesca Nerattini[1], Luca Tubiana[1], Chiara Cardelli[1], Valentino Bianco[1], Christoph Dellago[1] & Ivan Coluzza[2,3*]

Isolating the properties of proteins that allow them to convert sequence into the structure is a long-lasting biophysical problem. In particular, studies focused extensively on the effect of a reduced alphabet size on the folding properties. However, the natural alphabet is a compromise between versatility and optimisation of the available resources. Here, for the first time, we include the impact of the relative availability of the amino acids to extract from the 20 letters the core necessary for protein stability. We present a computational protein design scheme that involves the competition for resources between a protein and a potential interaction partner that, additionally, gives us the chance to investigate the effect of the reduced alphabet on protein-protein interactions. We devise a scheme that automatically identifies the optimal reduced set of letters for the design of the protein, and we observe that even alphabets reduced down to 4 letters allow for single protein folding. However, it is only with 6 letters that we achieve optimal folding, thus recovering experimental observations. Additionally, we notice that the binding between the protein and a potential interaction partner could not be avoided with the investigated reduced alphabets. Therefore, we suggest that aggregation could have been a driving force in the evolution of the large protein alphabet.

The amino acid alphabet encoding the protein function is common to all living organisms and is the result of millions of years of evolution. It is composed of 20 letters, in contrast to the ones of other biopolymers, such as DNA and RNA, which possess 4 letters only. Such a large alphabet gives to proteins the vast variety of configurations and functions that we know so far.

The advent of computational protein evolution (also known as protein design)[1–16] opens the possibility to address fundamental questions about the nature of the amino acid alphabet[17–20]. Protein design consists in searching for protein sequences capable of folding into a given backbone conformation. The search is usually done by point mutations while keeping the backbone structure fixed. In addition to several applications to medicine[12,14,21–23] and material science[15,24–27], protein design offers the possibility to explore fundamental problems of protein evolution.

One of the questions that mostly attracts the attention of the scientific community is about the universality of the 20 letters. Of course, the complex spectrum of proteins functionalities calls for a wide range of building blocks. However, could it be possible to design proteins to fold using a reduced alphabet? And, if yes, why not simply stick with such a reduced alphabet?

The early work on protein design with alphabets of different sizes was carried out for protein lattice models in which the protein chain is constrained to be on a cubic lattice. With such models it was possible to design heteropolymers with a large variety of alphabets defined by the amino acid interactions[28–37]. It became rapidly apparent that even in such simplified systems it is necessary to have a minimum number of residue types to encode the target configurations[38]. Moreover, such simple models allowed to explore the related question on how the alphabet size influences protein-protein interactions[39–42]. Finally, works done on realistic models offer substantial evidence that protein design with a minimalistic alphabet is possible[43–47]. In particular, statistical analysis of protein databases[48–54] demonstrated that a considerable fraction of the information encoded in natural proteins could be packed into smaller efficient alphabets from 12[54] all the way down to just 5 residue types[43,45,54–57]. However, all the

[1]Faculty of Physics, University of Vienna, Boltzmanngasse 5, 1090, Vienna, Austria. [2]Center for Cooperative Research in Biomaterials (CIC biomaGUNE), Basque Research and Technology Alliance (BRTA), Paseo Miramon 182, 20014, San Sebastian, Spain. [3]IKERBASQUE, Basque Foundation for Science, 48013, Bilbao, Spain. *email: icoluzza@cicbiomagune.es

mentioned studies completely neglect the possibility that a competition for the availability of amino acids may have played a role in the evolution of the protein alphabet size.

In this work, we devised a design strategy that includes such a competition to spontaneously drive the selection towards the minimal subset of residues essential for protein folding.

Our principal result is the identification of an optimal protein alphabet with the minimum number of letters, without the need of imposing neither the size nor the composition of it. The results show that for the folding of a small protein the minimum number of amino acid types needed is just 4. Incidentally, 4 is also the alphabet size of RNA that was hypothesized to be a precursor of proteins during the early stages of life. Additionally, by having a binary system, we can explore the effect of the alphabet reduction on aggregation in different protein-protein binding scenarios. From our simulations we observe that the alphabet reduction compromises the heterogeneity of the protein-protein interactions[28,36,40–42] and binding cannot be avoided.

These results have interesting implications towards the understanding of the evolution of protein sequences and structures when the amino acid availability is taken into account. In fact, living systems are under constant pressure for using the smallest variety of amino acids as possible, e.g. to limit the resources needed to construct specialised tRNA molecules necessary for the translation process[58]. Hence, it is reasonable to assume that during the early stages of life, the protein capable of being designed with a smaller alphabet could have been advantageous. If protein aggregation was not crucial at that stage, then our results demonstrate that protein-based life could have started with an alphabet size compatible with the one of DNA and RNA. On the other hand, the simple condition of avoiding protein aggregation could be a strong driving force against alphabet reduction.

Methods

We consider systems composed of the natural protein G structure (already successfully redesigned with several protein models[3,7]) and a competing element (a mould of a part of protein G, that mimics with a surface-like shape a potential binding site of a larger protein). Both proteins are represented with the caterpillar coarse-grain model, which has been successfully tested to design and refold natural and artificial proteins[7,9] including the protein G.

In the following we will use the denominations: protein G referring to both natural structure and sequence as stored in the PDB with the ID 1pgb; protein $\overline{G}$ referring to an artificial sequence designed for the natural protein G structure; protein Γ referring to the surface-like competing protein partner.

The protein Γ is created immersing the protein G structure into a flat surface until its centre of mass (CM) reaches the desired relative height ζ with respect to it. The flat surface is pushed down creating a mould, which is kept at fixed distance $\mu = 13$ Å from the surface of the protein G. Then, the protein G is rotated around its CM to maximise the mould surface area, which represents the binding site of a second protein. We create four moulds, each corresponding to a different value of ζ and composed by a different number of amino acids, labelled as C_{surf}. The systems are characterised by $\zeta = (0.20, 0.40, 0.60, 0.80)$, thus leading to surface areas = (4717.5, 3842.2, 3051.5, 2320.5) Å^2 and $C_{\text{surf}} = (158, 127, 100, 78)$ residues respectively (see the *Modelling protein* Γ of the Supplementary Materials SM for details). For the sake of simplicity, we call *sequence* the amino acid identities of protein Γ, although its surface-like structure is frozen and far from a polymeric chain of beads.

The procedure employed in the present work follows the steps pictorially represented in Fig. 1,

Once the protein Γ modelling is complete, the structures of both proteins are frozen, with the protein G immersed into the mould Γ and kept at distance μ from it (as represented in Fig. S1b). The design scheme consists of a computational exploration of the sequence space via point mutations, looking for the ones that minimise the total energy among the ones that maximise the permutations $N_P = \frac{N!}{n_A!\, n_B!\, n_C! \ldots}$ of the total amino acid composition (N is the total number of amino acids and $[n_A, n_B, n_C, \ldots]$ are the abundance on amino acids of type $A, B, C, \ldots$ respectively). See the subsection *Design* of *Technical aspects of the methodology* in the SM for details. It is important to stress that $N_P > N_P^{\overline{G}} N_P^{\Gamma}$, where $N_P^{\overline{G}}$ and N_P^{Γ} are the permutations of protein $\overline{G}$ and Γ respectively. This inequality implies that, indeed, the sequences of $\overline{G}$ and Γ are correlated, since the most heterogeneous sequence is not the one that maximises $N_P^{\overline{G}}$ and N_P^{Γ} separately. In turns it means also that N_P can be maximised without maximising $N_P^{\overline{G}}$ and N_P^{Γ} separately, and the residues can be distributed dishomogeneously between protein and substrate.

The choice of the distance μ between the two proteins guarantees that, during the design, the protein-protein interaction energy is negligible. Under such conditions, the design scheme leads inherently to sequences that minimise the energy of the protein $\overline{G}$ and optimise the exposure to the solvent of each residue of protein Γ. Since protein $\overline{G}$ and Γ are energetically uncorrelated, the coupling between the proteins is then only through the maximisation of the total permutations N_P.

Results

For each scenario, i.e. for each $\zeta \in (0.2, 0.4, 0.6, 0.8)$, the design algorithm generates a basin of solutions containing approximately 10^5 sequences. From each basin, we select the sequence with highest permutation number and lowest energy, considering it as representative of the whole basin, and use it to test the folding and binding properties. The selected protein $\overline{G}$ sequences for each scenario are shown in Table S1, while in Table S2 we show how much they differ from each other. To search for the smallest alphabet, we decided to focus on a single sequence instead of an average over a basin. Taking as a reference the centroid of the basin would have shifted the solution space towards higher energy sequences that tend to have larger alphabets.

We observe that the residues of protein $\overline{G}$ tend to adopt a limited set of letters. Moreover, increasing the protein Γ area (and hence the number of amino acids belonging to it) reduces de facto the amino acids accessible by

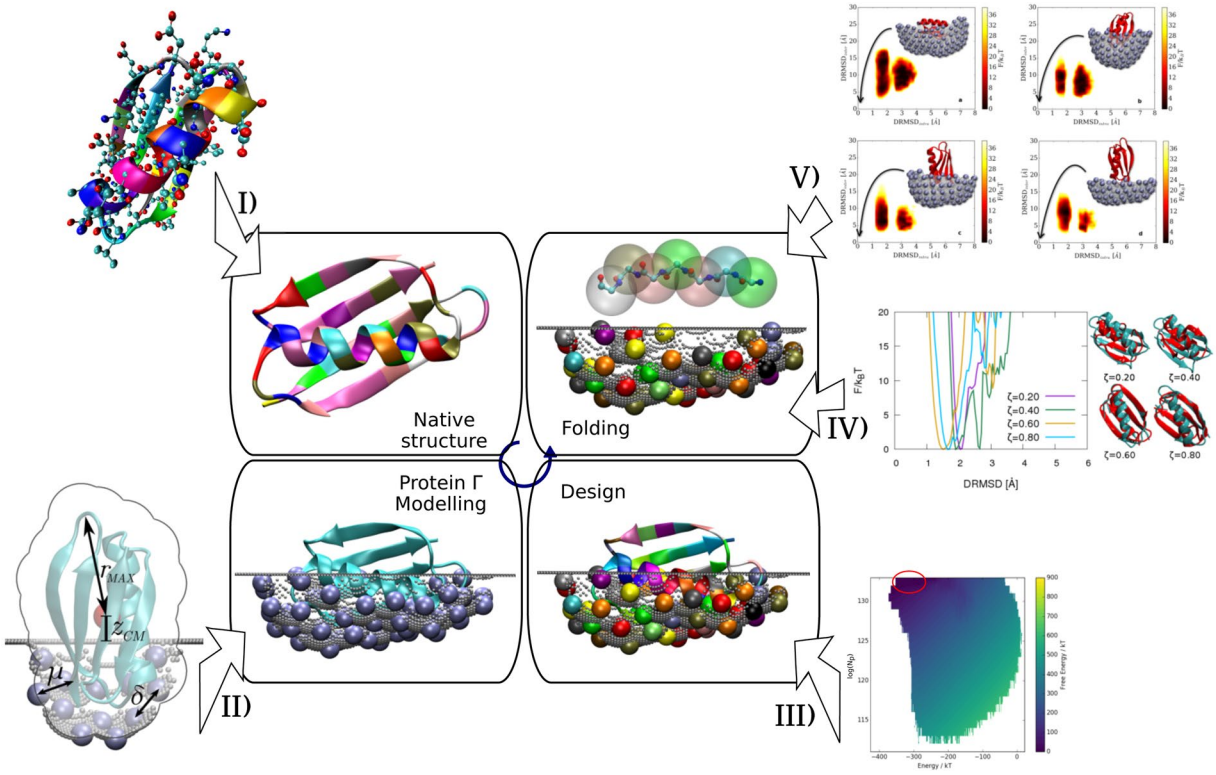

Figure 1. Pictorial representation of the steps employed to enforce a competition for amino acid availability between protein $\overline{G}$ and a protein Γ, and to test its effect on the folding ability of protein $\overline{G}$ in presence and absence of the artificial partner. (I) Create a Caterpillar version of the experimentally determined crystal structure of protein G (II) Shape four competing partner proteins Γ modelled as moulds of increasing portions of the protein G. The size of the mould will influence the competition for resources, as further explained in the following sections. The larger the surface, the higher the competition. (III) Design each of the four systems considering simultaneously the proteins $\overline{G}$ and Γ. The procedure consists in searching for the ensemble of sequences that minimise the energy of both protein $\overline{G}$ and Γ while keeping the system conformation frozen in space. The competition for the amino acids is created at this stage of our simulations. (IV) After selecting the best designed sequence (see the *Design* subsection for details about the criterion) for each system, isolate the portion relative to the protein $\overline{G}$ and test its folding ability in a single-protein folding simulation. (V) Check how the folding of the latter sequences is influenced by the presence of protein Γ frozen in the simulation box (bearing the sequence designed concurrently to protein $\overline{G}$).

the protein $\overline{G}$ to minimise its energy. Hence, the fractionation of the alphabet is not caused by specific interactions between the residues but by the coupling through the maximisation of the total permutations N_P.

We can control the competition pressure by changing the size of protein Γ. This competition leads to an effective reduced alphabet used by the protein $\overline{G}$. We observe that the effective alphabet grows from 4 to 6 letters going from larger ($\zeta = 0.20$ and 0.40) to smaller Γ proteins ($\zeta = 0.60$ and 0.80) respectively. It is interesting to notice that the alphabets are made of amino acids with an average attractive pair-interaction energy and high variability in terms of the residue-solvent interactions (see Table S1 in ref.[9]). Moreover, the alphabets differ from each other (letters *GKVY* and *GKRV* corresponding to $\zeta = (0.20, 0.40)$ and *FGHKRY* common to both $\zeta = (0.60, 0.80)$), and for each scenario the protein amino acids are not present in the corresponding protein Γ sequence (see SM Fig. S11). Therefore, part of the 20 letters are segregated on the protein Γ sequence.

Our finding shows that the design process indeed mimics a process under competition for available amino acids. It is important to stress that such competition is the results of the coupling alone as we impose neither the size nor the composition of the reduced alphabet. Hence, the particular letters that the design process chooses for protein $\overline{G}$ are presumably optimal to stabilise the folded structure. This feature is the crucial element of our design scheme that allows us to isolate the critical set of residues in our alphabet for design and folding.

Finally, we test the folding and binding properties of the designed sequences. Hence, we perform Monte Carlo simulations keeping fixed the amino acid sequence generated for each scenario, and extensively exploring the conformational space of the protein $\overline{G}$.

To test the selected sequences, we first examine the folding stability of the protein $\overline{G}$ alone, therefore performing a folding simulation in an empty box starting from a fully stretched configuration. Figure 2 shows the free energy profiles as a function of the distance root mean square displacement *DRMSD* (defined in Eq. S10 of SM). From previous works[7,9], the criterion for assessing a stable fold is to observe a funnel shape of the free energy

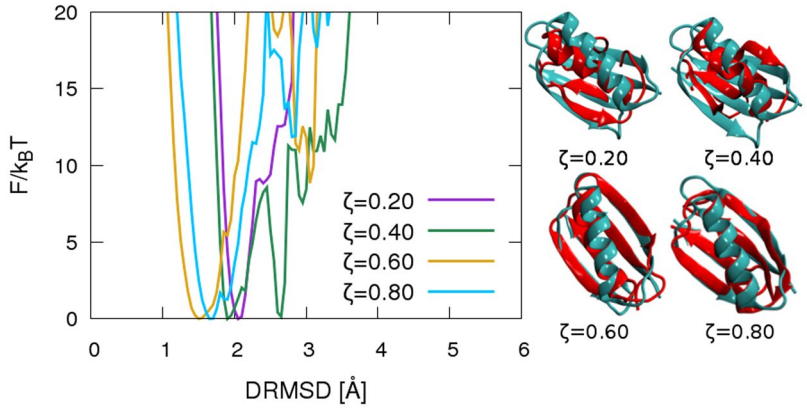

Figure 2. Folding free energy profiles F/k_BT of single protein (only protein $\overline{G}$, no protein Γ) at reduced temperature 0.55 as a function of DRMSD from the native target structure (protein G structure, PDB ID: 1pgb). Different colours correspond to protein $\overline{G}$ sequences obtained via the design procedure in the presence of the protein Γ characterised by the ζ value specified in the key. Right hand side: configurations corresponding to the free energy minimum for each system are represented in red, compared to the native protein G (in green). $DRMSD = 2.1$ Å for $\zeta = 0.20$; $DRMSD = 1.9$ Å for $\zeta = 0.40$; $DRMSD = 1.3$ Å for $\zeta = 0.60$ and $DRMSD = 1.5$ Å $\zeta = 0.80$.

profile and a global free energy minimum for $DRMSD \leq 2$ Å. Using this criterion, we can say that all protein sequences fold back into the target configuration, although with different precision. Sequences with a larger effective alphabet fold with higher precision, as can be seen from the *DRMSD* value of the configurations corresponding to the global free energy minimum for each system (The DRMSD values correspond to 4.9; 5.5; 2.4 and 2.7 Å in *RMSD* respectively). The sequence optimised at $\zeta = 0.40$ shows a secondary minimum in the free energy, corresponding to misfolded compact structures, therefore being the system less stable for the folding in the bulk. A possible explanation of such a behaviour is that the effective 4 letters protein $\overline{G}$ alphabet for $\zeta = 0.40$ involves only hydrophilic residues (*GKRY*), thus leading to a lower stability.

From the described scenario, we can draw two important conclusions: firstly, design with a limited alphabet of 4 letters can produce a funnel-like folding free energy landscape; secondly, with 6 letters we recover the folding precision of previous caterpillar designs made with 20 letters[9]. Our results are consistent with the experimental observation that 6 letters are a minimal set necessary to maintain protein structure and function[43,45,54–57].

The Random Energy Model[59–61] provides a criterion for a heteropolymer to be designable: it has to satisfy the relation $q > \exp(\omega)$, where q is the alphabet size and ω the conformational entropy per residue. Hence, a 4 letters alphabet gives an upper bound to the conformational entropy ω of the caterpillar backbone and therefore of the more restricted natural protein backbones. Such a result is compatible with the recent observations of Cardelli *et al.*[62] who mapped the designability phase space for a general heteropolymer decorated with directional interactions similar to the hydrogen bonds present along the protein backbone. For polymers with two directional interactions per particle the minimum alphabet measured was four, as the one presented here.

To test the effect of the alphabet reduction on protein-protein interaction, we also perform folding simulations in the presence of the protein Γ, that represent a potential binding site. In Fig. 3 we plot the free energy landscape as a function of $DRMSD_{intra}$ and $DRMSD_{inter}$. $DRMSD_{intra}$ is the *DRMSD* intra protein $\overline{G}$, and uses the native protein G structure as target configuration. $DRMSD_{inter}$ is the *DRMSD* between protein $\overline{G}$ and protein Γ, and uses the folded bound configuration (shown in the insets of Fig. 3 for each scenario) as a target. This choice allows us to monitor the folding and binding properties of the system independently. Conformations that are folded and bound can be found in the bottom left corner, while folded unbound ones in the top left corner.

Additionally, we also separately check the free energy profiles as a function of $DRMSD_{intra}$ for conformations with protein $\overline{G}$ in contact with protein Γ (see Fig. S9 in the SM) and in the bulk solution (i.e. where no inter-protein contacts are possible, see Fig. S10) in the SM. For a sketch of the definition of contact and bulk solution configurations see Fig. S6 in the SM. To verify the consistency of the two different folding simulations, we checked that the free energy profiles of configurations in the latter region correctly fold into the target structure (Fig. S10), reproducing the behaviour observed in the isolated protein folding simulations (Fig. 2).

For all scenarios, upon binding to protein Γ, we observe a significant enhancement of misfolded configurations with respect to what observed in the bulk solution (compare Figs. S9(a) and S10 in the SM). In particular, there is a considerable shift in the equilibrium towards states at $DRMSD \sim 3$ Å that have a free energy that is now comparable to the one of properly folded configurations. It should be noticed that natural binding sites expose much smaller surface areas then the one modelled with protein Γ. Hence, the latter effect might be mitigated considering smaller surfaces for protein Γ.

Analysing the behaviour of the binding process as a function of temperature we find that the random binding is overall very strong and it decreases while increasing the temperature. The van't Hoff plot[63,64] shows positive binding affinities and an exothermic process above the folding temperature (Fig. S7 SM; see Fig. S6 SM for details

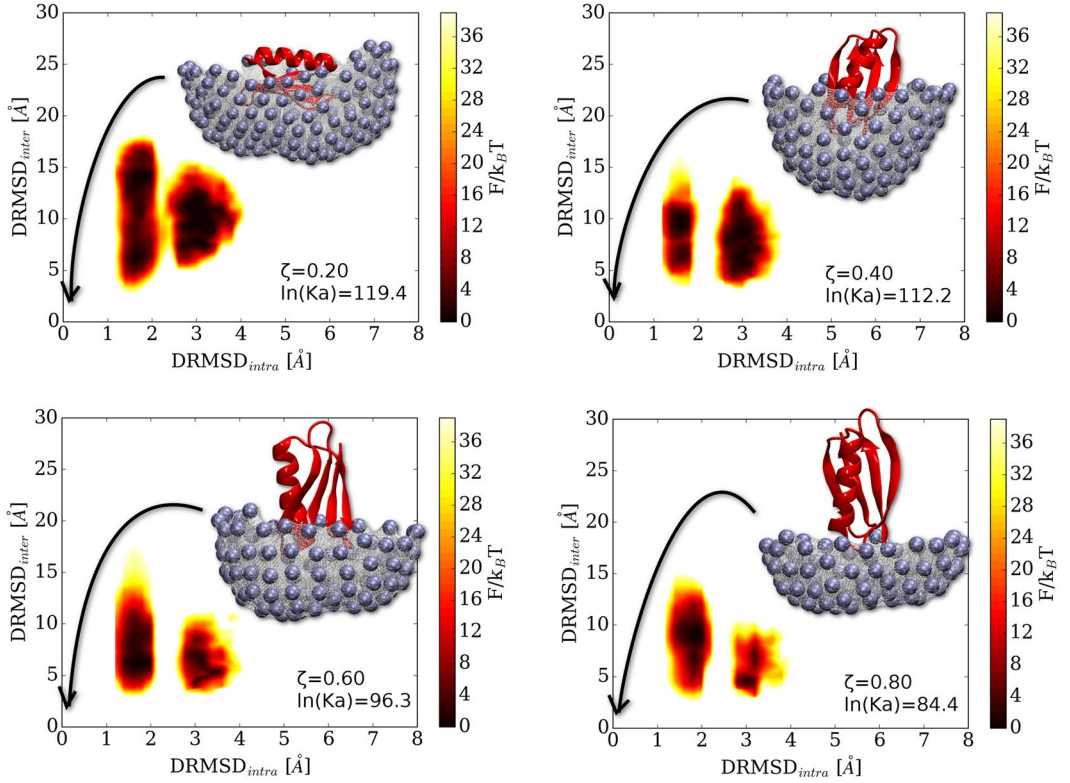

Figure 3. Folding free energy landscapes F/k_BT at reduced temperature 0.76 as a function of the $DRMSD_{intra}$ distance from the native protein G as target and the $DRMSD_{inter}$ inter-protein distance from the folded protein bound to protein Γ (configurations depicted in the panels). The binding affinity decreases along with the protein Γ surface size, as shown by the value of the association constants K_a in the plot key.

about the evaluation of the association constant and Fig. S8 SM for the folding temperature evaluation). At the same time, while increasing the temperature, the equilibrium shifts from partially-misfolded to fully-misfolded, indicating that the unfolding process takes place at the surface while the protein remains bound (see Fig. S9(b)). This is particularly evident for extended protein Γ surfaces, i.e. systems characterised by $\zeta = (0.20,\ 0.40)$. Hence, we observe a strong tendency of the protein $\overline{G}$ designed with 4 letters to absorb and aggregate on protein Γ.

Overall, this binding behaviour is an unexpected result. In the crowded cellular ambient, natural protein designed by evolution with the 20 letters alphabet are not aggregating. As such, in the present work, protein $\overline{G}$ and Γ should not aggregate, since they interact through the total caterpillar alphabet of 18 letters. However, our design scheme imposes a segregation of few letters on the protein $\overline{G}$ sequence. We identify the following observation as a possible cause. The 4 letters alphabets (*GKVY* and *GKRY* corresponding to $\zeta = (0.20,\ 0.40)$) have an average intra-protein residue interaction of $-0.2k_BT$, while the average interaction of the single protein $\overline{G}$ letters with all the others, i.e. the inter-protein interaction, is much lower $-0.3k_BT$. This makes impossible for the protein $\overline{G}$ to stabilise the folded state in contact with protein Γ. Conversely, the 6 letter alphabet (*FGHKRY* common to both $\zeta = (0.60,\ 0.80)$) has an average intra-protein residue interaction of $-0.4k_BT$, that is lower than the inter-protein one of $-0.3k_BT$. This helps in stabilizing the folded structure upon binding. If, on the other end, the residues would have been properly mixed, there would be no difference between inter and intra averages, and the random interactions should be washed out by thermal fluctuations[28]. Hence, there is a fundamental pressure to increase the alphabet size and fully use it to achieve folding and avoid strong absorption.

This is an essential factor that could explain why natural proteins tend to have and use a larger alphabet than 6 letters. However, the origin of the 20 letters is still only matter of speculation. In fact many molecular process require additional chemical modification of the proteins like glycolisation that effectively increases the available pools of potential letters. Hence, it is not even accurate to consider 20 as the upper limit, that is why in this study we focused on the lower limit that has more clear definition.

In conclusion, the design procedure employed in our work has a significant segregation effect on the alphabet letters used in the protein $\overline{G}$ sequence. The larger the number of residues on the competing protein Γ, the smaller is the effective alphabet available for the protein $\overline{G}$ sequence. On the one side, the design is capable of selecting a subset of letters that still allows the folding of the protein in the bulk solution even for the smallest effective alphabet (4 letters). The precision of the folding increases with the effective alphabet size. Interestingly, the experimentally determined minimum alphabet size of 6 letters is also what we identify as minimum alphabet that recovers the design accuracy commonly obtained with a 20 letter alphabet. This implies that functionality will push the

alphabet to grow. This trend could explain why reduced alphabets obtained form the analysis of natural proteins then to be larger[54].

It is important to stress that the reduced alphabet presented here might not be the only possible solution. It would be interesting to perform a larger study of the folding sequences and generate a spectrum of possible 4 letters alphabets, and with models that include amino acids charges more explicitly.

Our results have far-reaching implications both in the field of protein design and for the understanding of protein evolution. In protein design, the possibility of using a reduced alphabet would considerably accelerate the search of the sequence space for good folders. In the field of protein evolution instead, the understanding of the smallest alphabet necessary for accurate protein design is still an open question. To the best of our knowledge, this study represents the first successful design of a full natural protein structure with a reduced alphabet of just 4 letters. Moreover, such a result offers an interesting parallelism with the 4 letter alphabet of RNA which studies speculates had a role in the early stages of life before the advent of proteins.

Received: 27 February 2019; Accepted: 8 December 2019;
Published online: 14 February 2020

References

1. Gutin, A. M. & Shakhnovich, E. Ground-state of random copolymers and the discrete Random Energy-model. *J. Chem. Phys.* **98**, 8174–8177, https://doi.org/10.1063/1.464522 (1993).
2. Dahiyat, B. I. & Mayo, S. De Novo Protein Design: Fully Automated Sequence Selection. *Sci. (80-).* **278**, 82–87, https://doi.org/10.1126/science.278.5335.82 (1997).
3. Koehl, P. & Levitt, M. De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161–81, https://doi.org/10.1006/jmbi.1999.3211 (1999).
4. Kortemme, T. & Baker, D. Computational design of protein-protein interactions. *Curr. Opin. Chem. Biol.* **8**, 91–97, https://doi.org/10.1016/j.cbpa.2003.12.008 (2004).
5. Fung, H. K., Welsh, W. J. & Floudas, C. A. Computational de novo peptide and protein design: Rigid templates versus flexible templates. *Ind. Eng. Chem. Res.* **47**, 993–1001, https://doi.org/10.1021/ie071286k (2008).
6. Samish, I., Macdermaid, C., Perez-Aguilar, J. & Saven, J. Theoretical and computational protein design. *Annu. Rev. Phys. Chem.* **62**, 129–149, https://doi.org/10.1146/annurev-physchem-032210-103509 (2011).
7. Coluzza, I. A coarse-grained approach to protein design: learning from design to understand folding. *Plos One* **6**, e20853, https://doi.org/10.1371/journal.pone.0020853 (2011).
8. Koga, N. *et al*. Principles for designing ideal protein structures. *Nature* **491**, 222–227, https://doi.org/10.1038/nature11600, NIHMS150003 (2012).
9. Coluzza, I. Transferable Coarse-Grained Potential for De Novo Protein Folding and Design. *Plos One* **9**, e112852, https://doi.org/10.1371/journal.pone.0112852, arXiv:1406.4373v1 (2014).
10. Thomson, A. R. *et al*. Computational design of water-soluble a-helical barrels. *Sci. (80-).* **346**, 485–488, https://doi.org/10.1126/science.1257452 (2014).
11. Sevy, A. M., Jacobs, T. M., Crowe, J. E. & Meiler, J. Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *Plos Comput. Biol.* **11**, e1004300, https://doi.org/10.1371/journal.pcbi.1004300 (2015).
12. Pelay-Gimeno, M., Glas, A., Koch, O. & Grossmann, T. N. Structure-Based Design of Inhibitors of Protein-Protein Interactions: Mimicking Peptide Binding Epitopes. *Angew. Chemie - Int. Ed.* **54**, 8896–8927, https://doi.org/10.1002/anie.201412070 (2015).
13. Chevalier, A. *et al*. Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79, https://doi.org/10.1038/nature23912 (2017).
14. Marcos, E. *et al*. Principles for designing proteins with cavities formed by curved b sheets. *Sci. (80-).* **355**, 201–206, https://doi.org/10.1126/science.aah7389 (2017).
15. Coluzza, I. *et al*. Perspectives on the future of ice nucleation research: Research needs and Unanswered questions identified from two international workshops. *Atmosphere* (*Basel*). **8**, https://doi.org/10.3390/atmos8080138 (2017).
16. Bianco, V., Pagès-Gelabert, N., Coluzza, I. & Franzese, G. How the stability of a folded protein depends on interfacial water properties and residue-residue interactions. *J. Mol. Liq.* **245**, 129–139 (2017).
17. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci.* **91**, 2146–2150, https://doi.org/10.1073/pnas.91.6.2146 (1994).
18. Riddle, D. S. *et al*. Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809, https://doi.org/10.1038/nsb1097-805 (1997).
19. Cordes, M. H. J., Davidson, A. R. & Sauer, R. T. Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3–10, https://doi.org/10.1016/S0959-440X(96)80088-1 (1996).
20. Davidson, A. R., Lumb, K. J. & Sauer, R. T. Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* **2**, 856 (1995).
21. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327, https://doi.org/10.1038/nature19946 (2016).
22. Parmeggiani, F. & Huang, P.-S. Designing repeat proteins: a modular approach to protein design. *Curr. Opin. Struct. Biol.* **45**, 116–123, https://doi.org/10.1016/j.sbi.2017.02.001 (2017).
23. Baran, D. *et al*. Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci.* **114**, 10900–10905, https://doi.org/10.1073/pnas.1707171114 (2017).
24. Mejias, S. H. *et al*. Repeat protein scaffolds: ordering photo- and electroactive molecules in solution and solid state. *Chem. Sci.* **7**, 4842–4847, https://doi.org/10.1039/C6SC01306F (2016).
25. Cortajarena, A. L., Liu, T. Y., Hochstrasser, M. & Regan, L. Designed Proteins To Modulate Cellular Networks. *ACS Chem. Biol.* **5**, 545–552, https://doi.org/10.1021/cb9002464 (2010).
26. Mejias, S. H., Aires, A., Couleaud, P. & Cortajarena, A. L. Designed Repeat Proteins as Building Blocks for Nanofabrication. In Cortajarena, A. L. & Grove, T. Z. (eds) *Adv. Exp. Med. Biol.*, vol. 940, chap. Protein-ba, 61–81, https://doi.org/10.1007/978-3-319-39196-0_4 (Springer International Publishing, Cham, 2016).
27. Bianchi, E., Capone, B., Coluzza, I., Rovigatti, L. & van Oostrum, P. D. J. Limiting the valence: advancements and new perspectives on patchy colloids, soft functionalized nanoparticles and biomolecules. *Phys. Chem. Chem. Phys.* **19**, 19847–19868, https://doi.org/10.1039/C7CP03149A, 1705.04383 (2017).
28. Coluzza, I. & Frenkel, D. Designing specificity of protein-substrate interactions. *Phys. Rev. E* **70**, 51917, https://doi.org/10.1103/PhysRevE.70.051917 (2004).
29. Coluzza, I., Muller, H. G. & Frenkel, D. Designing refoldable model molecules. *Phys. Rev. E* **68**, 046703, https://doi.org/10.1103/PhysRevE.68.046703 (2003).

30. Salvi, G., Mölbert, S. & De Los Rios, P. Design of lattice proteins with explicit solvent. *Phys. Rev. E* **66**, 61911, https://doi.org/10.1103/PhysRevE.66.061911 (2002).
31. Wang, T. R., Miller, J., Wingreen, N. S., Tang, C. & Dill, K. A. Symmetry and designability for lattice protein models. *J. Chem. Phys.* **113**, 8329–8336, https://doi.org/10.1063/1.1315324, 0006372 (2000).
32. Deutsch, J. M. & Kurosky, T. A New Algorithm for Protein Design. *Phys. Rev. Lett.* **76**, 10, https://doi.org/10.1103/PhysRevLett.76.323, 9508127 (1995).
33. Shakhnovich, E. I. & Gutin, A. M. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci.* **90**, 7195–7199, https://doi.org/10.1073/pnas.90.15.7195 (1993).
34. Yue, K. & Dill, K. A. Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. USA* **89**, 4163–4167, https://doi.org/10.1073/pnas.89.9.4163 (1992).
35. Bryngelson, J. D. D. & Wolynes, P. G. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528, https://doi.org/10.1073/pnas.84.21.7524 (1987).
36. Coluzza, I. & Frenkel, D. Monte Carlo study of substrate-induced folding and refolding of lattice proteins. *Biophys. J.* **92**, 1150–1156, https://doi.org/10.1529/biophysj.106.084236 (2007).
37. Abeln, S. & Frenkel, D. Disordered Flanks Prevent Peptide Aggregation. *Plos Comput. Biol.* **4**, e1000241, https://doi.org/10.1371/journal.pcbi.1000241 (2008).
38. Chan, H. S. & Dill, K. A. Comparing folding codes for proteins and polymers. *Proteins Struct. Funct. Genet.* **24**, 335–344, https://doi.org/10.1002/(SICI)1097-0134(199603)24:3h335::AID-PROT6i3.0.CO;2-F (1996).
39. Sear, R. P. & Cuesta, J. A. Instabilities in Complex Mixtures with a Large Number of Components. *Phys. Rev. Lett.* **91**, 245701, https://doi.org/10.1103/PhysRevLett.91.245701, 0307326 (2003).
40. Sear, R. P. Specific protein–protein binding in many-component mixtures of proteins. *Phys. Biol.* **1**, 53–60, https://doi.org/10.1088/1478-3967/1/2/001, 0312033 (2004).
41. Sear, R. P. Highly specific protein–protein interactions, evolution and negative design. *Phys. Biol.* **1**, 166–172, https://doi.org/10.1088/1478-3967/1/3/004 (2004).
42. Madge, J. & Miller, M. A. Design strategies for self-assembly of discrete targets. *J. Chem. Phys.* **143**, 044905, https://doi.org/10.1063/1.4927671 (2015).
43. Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. Simplified proteins: Minimalist solutions to the 'protein folding problem'. *Curr. Opin. Struct. Biol.* **8**, 80–85, https://doi.org/10.1016/S0959-440X(98)80013-4 (1998).
44. Walter, K. U., Vamvaca, K. & Hilvert, D. An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.* **280**, 37742–37746, https://doi.org/10.1074/jbc.M507210200, jbc.M507210200 (2005).
45. Reetz, M. T. & Wu, S. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun.* **5499**, https://doi.org/10.1039/b813388c (2008).
46. Liu, B. *et al.* IDNA-Prot—dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *Plos One* **9**, https://doi.org/10.1371/journal.pone.0106691 (2014).
47. Sun, Z. *et al.* Reshaping an Enzyme Binding Pocket for Enhanced and Inverted Stereoselectivity: Use of Smallest Amino Acid Alphabets in Directed. *Evolution. Angew. Chemie - Int. Ed.* **54**, 12410–12415, https://doi.org/10.1002/anie.201501809 (2015).
48. Wang, J. & Wang, W. Simplification of complexity in protein molecular systems by grouping amino acids: a view from physics. *Adv. Phys. X* **1**, 444–466, https://doi.org/10.1080/23746149.2016.1216329 (2016).
49. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2014).
50. Ferreiro, D. U., Komives, E. A. & Wolynes, P. G. Frustration in biomolecules. *Q. Rev. Biophys.* **47**, 285–363, https://doi.org/10.1017/S0033583514000092 (2014).
51. Uversky, V. N. A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Sci.* **22**, 693–724, https://doi.org/10.1002/pro.2261 (2013).
52. Longo, L. M. & Blaber, M. Protein design at the interface of the pre-biotic and biotic worlds. *Arch. Biochem. Biophys.* **526**, 16–21, https://doi.org/10.1016/j.abb.2012.06.009 (2012).
53. Li, T., Fan, K., Wang, J. & Wang, W. Reduction of protein sequence complexity by residue grouping. *Protein Eng.* **16**, 323–330, https://doi.org/10.1093/protein/gzg044 (2003).
54. Murphy, L. R., Wallqvist, A. & Levy, R. M. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* **13**, 149–152, https://doi.org/10.1093/protein/13.3.149 (2000).
55. Chan, H. S. Folding alphabets. *Nat. Struct. Biol.* **6**, 994–6, https://doi.org/10.1038/14876 (1999).
56. Wang, J. & Wang, W. A computational approach to simplifying the protein folding alphabet. *Nat. Struct. Biol.* **6**, 1033–1038, https://doi.org/10.1038/14918 (1999).
57. Solis, A. D. Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. *Proteins Struct. Funct. Bioinforma.* **83**, 2198–2216, https://doi.org/10.1002/prot.24936 (2015).
58. Alberts, B. *et al. Molecular Biology of the Cell* (Garland Science, 2002).
59. Derrida, B. Phenomenological Renormalization Of The Self Avoiding Walk In 2 Dimensions. *J. Phys. A-Mathematical Gen.* **14**, L5–L9 (1981).
60. Pande, V. S., Grosberg, A. Y. & Tanaka, T. Heteropolymer freezing and design: Towards physical models of protein folding. *Rev. Mod. Phys.* **72**, 259–314, https://doi.org/10.1103/RevModPhys.72.259 (2000).
61. Pande, V. S. V., Grosberg, A. Y. A. & Tanaka, T. Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192–3210, https://doi.org/10.1016/S0006-3495(97)78345-0 (1997).
62. Cardelli, C. *et al.* The role of directional interactions in the designability of generalized heteropolymers. *Sci. Rep.* **7**, 4986, https://doi.org/10.1038/s41598-017-04720-7 (2017).
63. Lim, C. W. & Kim, T. W. Dynamic [2]Catenation of Pd(II) Self-assembled Macrocycles in Water. *Chem. Lett.* **41**, 70–72, https://doi.org/10.1246/cl.2012.70 (2012).
64. Hino, S., Ichikawa, T. & Kojima, Y. Thermodynamic properties of metal amides determined by ammonia pressurecomposition isotherms. *J. Chem. Thermodyn.* **42**, 140–143, https://doi.org/10.1016/j.jct.2009.07.024 (2010).

Acknowledgements

All simulations presented in this paper were carried out on the Vienna Scientific Cluster (VSC). We acknowledge support from the VSC School, as well as from the Austrian Science Fund (FWF) project 26253-N27. V.B. acknowledges the support from FWF Grant No. M 2150-N36. I.C. gratefully acknowledges support from the Ministerio de Economià y Competitividad (MINECO) (FIS2017-89471-R). This work was performed under the Maria de Maeztu Units of Excellence Program from the Spanish State Research Agency – Grant No. MDM-2017-0720.

Author contributions

I.C. designed the research, F.N. performed the simulations, F.N. and I.C. performed the data analysis. C.C., F.N., V.B., L.T., I.C. and C.D. wrote the manuscript and discussed the research.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-59401-9.

Correspondence and requests for materials should be addressed to I.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.