molecules, it should now be possible to generate large numbers of aptazymes against large numbers of effectors. Indeed, aptamers have been selected against targets that range from zinc to whole organisms[13,14] and there is no reason to believe that the recognition abilities of aptazymes will be any less.

The transformation of molecular recognition directly to molecular catalysis affords a variety of opportunities for signal generation. For example, hammerhead aptazymes could potentially cleave quenchers away from fluors. Similarly, ligase aptazymes could potentially co-immobilize oligonucleotides bearing a variety of reporters, from fluors to enzymes to magnetic particles. Interestingly, aptazyme ligases have the unique property of being able to transduce effectors into templates that can be amplified, affording an additional boost in signal prior to detection[4]. Aptazymes can be mounted in arrays using some of the same technologies that have been used to create 'DNA chips' but could instead be used to monitor the presence and concentrations of different metabolites or proteins, rather than mRNAs (Fig. 3). Finally, the new frontier for aptazymes will be *in vivo*, where they can potentially be used to develop programmed genetic circuits that can be used to regulate gene expression. In particular, aptazymes may abet the development of exquisitely regulated gene therapies.

Kristin A. Marshall and Andrew D. Ellington are in the Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712, USA.

Correspondence should be addressed to A.D.E. email: andy.ellington@mail.utexas.edu

1. Pace, N.R. & Marsh, T.L. *Orig. Life Evol. Biosph.* **16**, 97-116 (1985).
2. Torres, R.A. & Bruice, T.C. *Proc. Natl. Acad. Sci. USA* **95**, 11077-11082 (1998).
3. Koizumi, M., Soukup, G.A., Kerr, J.N.Q. & Breaker, R.R. *Nature Struct. Biol.* **6**, 1062–1071 (1999).
4. Robertson, M.P. & Ellington, A.D. *Nature Biotechnol.* **17**, 62-66 (1999).
5. Tang, J. & Breaker, R.R. *Chem. Biol.* **4**, 453-459 (1997).
6. Jiang, F., Kumar, R.A., Jones, R.A. & Patel, D.J. *Nature* **382**, 183-186 (1996).
7. Tang, J. & Breaker, R.R. *Nucleic Acids Res.* **26**, 4214-4221 (1998).
8. Soukup, G.A. & Breaker, R.R. *Structure Fold. Des.* **7**, 783-791 (1999).
9. Soukup, G.A. & Breaker, R.R. *Proc. Natl. Acad. Sci. USA* **96**, 3584-3589 (1999).
10. Hesselberth, J., Robertson, M.P., Jhaveri, S. & Ellington, A.D. *Reviews in Molecular Biotechnology*, **in the press** (1999).
11. Lorsch, J.R. & Szostak, J.W. *Nature* **371**, 31-6 (1994).
12. Ellington, A.D. & Robertson, M.P. In *Comprehensive natural products chemistry*. (eds. Soll, D., Nishimura, S. & Moore, P.B.) 115–148 (Elsevier, Oxford; 1999).
13. Jayasena, S.D. *Clin. Chem.* **45**, 1628-1650 (1999).
14. Famulok, M. & Jenne, A. *Curr. Opin. Chem. Biol.* **2**, 320-327 (1998).

Folding alphabets

Hue Sun Chan

A new computational approach optimizes searches for reduced protein folding alphabets that use fewer than 20 types of amino acids. The predicted optimal five-letter alphabet happens to be in agreement with the suggestive results of a recent experiment, but whether highly reduced alphabets are sufficient for truly protein-like properties remains an open experimental question.

The prospect of using a reduced alphabet to achieve protein-like properties is appealing. There are a number of reasons for this, not the least of which is an intriguing suggestion that primordial forms of life might have once operated on a reduced alphabet[1]. In addition, it has been thought that for polypeptide chains consisting of fewer than 20 letters (that is, the 20 types of common amino acids), the physics and chemistry may be sufficiently simplified for a thorough understanding of the protein folding code. However, as has been emphasized by recent theoretical considerations (reviewed in ref. 2), a certain threshold of heterogeneity or diversity in interaction energies must be present for the polypeptides to have protein-like properties. This requirement is intuitive, as one-letter homopolymers[3] do not have unique native structures like proteins. Experimental choices of reduced alphabets, with three or more letters, have been made with this general criterion in mind. Is it possible to be more systematic in selecting reduced alphabets? On page 1033 of this issue of *Nature Structural Biology*, Wang and Wang[4] propose an algorithm for this purpose.

To arrive at an optimized reduced alphabet of size N, Wang and Wang's program of reduction first divides the 20 amino acid types into N groups, with the aim of picking a representative from each group for the final reduced alphabet. In the end, all possible groupings with N groups are considered. The best grouping is selected by a 'minimal mismatch' principle, which ensures that all interactions between amino acids belonging to any two given groups are as similar to one another as possible. If amino acids with very different properties were placed together within one group, the resulting groupings would have high levels of mismatch. Such groupings are undesirable because a significant amount of interaction heterogeneity would be lost when only one amino acid is selected from each group for the final alphabet. Wang and Wang's procedure disfavors such occurrences by minimizing mismatch. The main tenets of their proposal are given in Fig. 1. Reduction of the 20-letter alphabet entails more involved combinatorial procedures and extensive Monte Carlo sampling[4], but the principles are basically the same. One of the optimally reduced sets predicted by Wang and Wang's efforts is a five-letter alphabet: Ile, Ala, Gly, Glu, Lys (IAGEK).

Experimental interest in reduced alphabets has existed for decades. A study in 1967 found many α-helices in soluble random polypeptides of Ala, Glu, and Lys (AEK)[5]. Another study by Rao *et al.*[6] in 1974 indicated that random AEK sequences can collapse to compact, globular conformations with helical content of 46% or higher, although no individual sequence was identified. Recent years have witnessed an increase in research efforts based on the reduced-alphabet theme; the following are some examples.

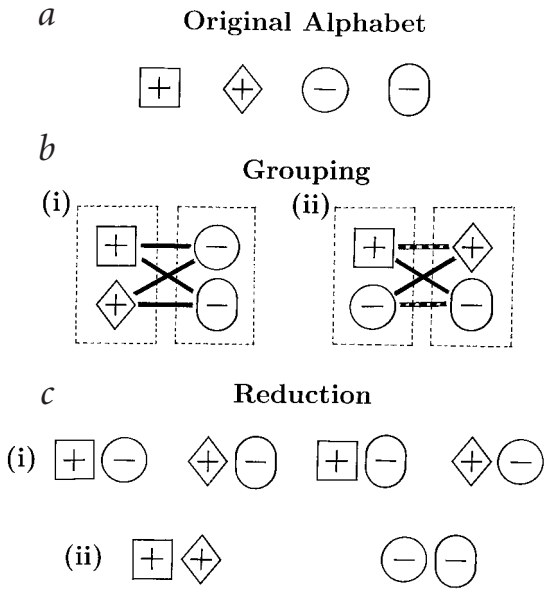

Fig. 1 An illustration of how optimized reduced alphabets are deduced by the mismatch minimization procedure of Wang and Wang. ***a***, A hypothetical alphabet with four residue types. In this simple example, two residue types are positively charged and the other two are negatively charged. As usual, opposite charges attract (favorable interactions) and like charges repel (unfavorable interactions). The principles are the same with the addition of more residues and interaction types, such as hydrophobic and polar interactions. ***b***, Let our goal be a reduced alphabet with two residue types ($N = 2$). Dividing the four residue types into two groups can be done with one residue in one group and three residues in the other group, {1,3}, or with two residues in both groups, {2,2}. There are six possible groupings for {2,2}, two of which are shown here (dotted boxes). The mismatches in interactions are then determined for all possible groupings. Grouping (i) has little mismatch because all individual interactions between elements of the two groups are favorable (solid lines). On the other hand, grouping (ii) has a high level of mismatch, or inconsistency, because some interactions are favorable (solid lines) while others are unfavorable (broken lines). Therefore, by minimizing mismatch, the procedure of Wang and Wang will choose grouping (i) in favor of grouping (ii). ***c***, In the final step of reduction, one of the residues in each group is chosen as the representative of that group. This is determined by another mismatch minimization. The set of all representative residues constitutes the reduced alphabet. There are six possible two-letter reduced alphabets in this example. The reduced alphabets in (i) contain both positively and negatively charged residues; thus, they preserve the possibility of having favorable as well as unfavorable interactions in a protein chain. However, the reduced alphabets in (ii) contain only positively charged or only negatively charged residues; hence, it becomes impossible to have favorable interactions in these two reduced alphabets. The reductions in (ii) are therefore undesirable; they represent a serious distortion of the physics allowed by the original four-letter alphabet in (*a*), as interaction heterogeneity is drastically diminished. However, these unphysical reductions cannot arise in Wang and Wang's procedure; they could only emerge from groupings with serious mismatches, such as (ii) in part (*b*), but these groupings would have been discarded by mismatch minimization of interaction energies in the first step of the procedure.

Munson *et al.*[7] successfully reconstituted the 32-residue hydrophobic core of the homodimer four-helix-bundle protein Rop using only Ala and Leu residues, while maintaining the native structure and physical properties of the wild type. Rojas *et al.*[8] designed *de novo* sequences based on hydrophobic-polar binary patterns for four-helix-bundle folds, a significant fraction of which were found to be capable of binding heme. Riddle *et al.*[9] found that functional SH3 domains can be encoded using only five types of amino acid for a majority of the chain (Ile, Ala, Gly, Glu, Lys; IAGEK). Brown and Sauer[10] synthesized a mutant of phage P22 Arc repressor with 15 non-alanine residues (of a total of 53 positions) mutated to Ala, and it still exhibited native-like properties.

These and similar studies (reviewed in ref. 11) have provided valuable insight. Indeed, the five-letter alphabet of Baker's group[9] and a theoretical perspective of Wolynes[2] were the main incentives for Wang and Wang's analysis[4]. By themselves, however, these experiments are insufficient for a definitive conclusion regarding the viability of significantly reduced alphabets. The reason is that, notwithstanding the elegant simplifications involved in their design, the numbers of amino acid types used to encode the protein-like sequences in all these cases are actually close to 20. The presence of many amino acid types arises either from variations intrinsic to a particular sequence construction method[8] or from variations at positions not restricted to a reduced alphabet in the experimental design[7,9,10]. In the four recent studies cited above, the most simplified sequences reported contain 16, 14, 14, and 16 amino acid types, respectively. For example, 16 positions (29%) of the simplest sequence studied by Riddle *et al.*[9] are not encoded by the IAGEK alphabet. Therefore, while highly suggestive, the exact agreement between the five-letter alphabet (IAGEK) of Baker's group and the one optimized by Wang and Wang (IAGEK) may be only coincidental.

Instead of targeting sequences within the context of a particular structure, a complementary line of inquiry is constructing random sequence libraries using reduced alphabets, in the spirit of the early AEK 'scattergun' approach of Rao *et al.*[6], with no input of structural information. A few years ago, Davidson *et al.*[12] found that 1–2% of random sequences from a predominantly three-letter Glu, Leu, Arg (QLR) alphabet possessed native-like protein properties, such as high helical content, cooperative unfolding, and discrete oligomeric states. However, in subsequent experiments on random sequence libraries composed of a three-letter Asp, Ile, Lys (NIK) alphabet or an essentially full (16-letter) alphabet, proteins with native properties similar to those of the QLR proteins were not found at a high enough frequency to be detected by the screening procedure employed (frequency < 0.1%) (A.R. Davidson, M.H.J Cordes, and R.T. Sauer, pers. comm.). The latter result is consistent with the recent suggestion by Yamauchi *et al.*[13] that it may be easier for certain reduced alphabets such as QLR[10] (and AEK[6]) to achieve high secondary structure content than 20-letter random polypeptides. In any event, even the folded structures of the 'protein-like' QLR sequences do not pack as well as natural proteins — instead they resemble molten globules[12]. It should be noted that QLR sequences do contain other amino acids (~14%), although most of them are at the chain ends and were not included deliberately for structural purposes.

How much can the amino acid alphabet be reduced? Apparently, experimental evidence to date indicates that it is possible to achieve native protein properties with seven-letter alphabets[11]. Examples include an AEKDSTR alphabet that encodes a number of natural helical antifreeze proteins in fishes[14], and a QLAGEKS alphabet used in a designed sequence that folds to a four-helix bundle[15]. It is not yet clear, however, how much structural and functional diversity these reduced alphabets can support.

Wang and Wang's reduced alphabets are theoretical. Their analysis is based on a residue–residue statistical potential derived by Miyazawa and Jernigan[16] (and therefore denoted the MJ potential) from a native structure database[17,18]. This is a 'knowledge-based' potential, consisting of interaction energies for all possible pairs of amino acids. These energies are deduced from the frequencies of contacts between different amino acid residues in a set of known protein structures. The MJ and similar interaction parameters have

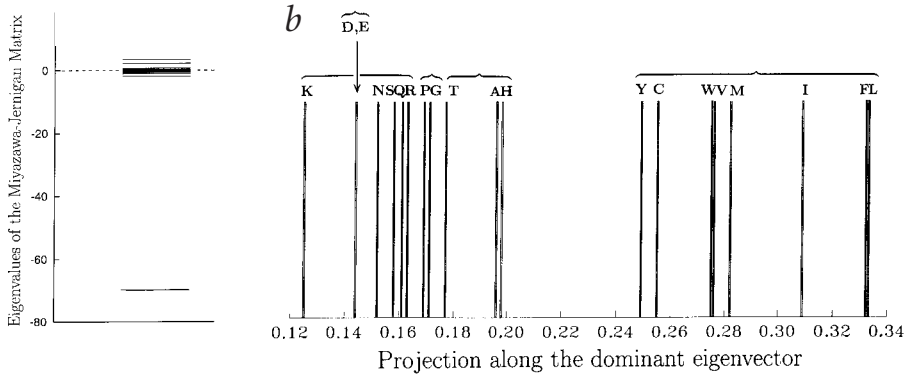

Fig. 2 Reduced alphabets by mismatch minimization follow an approximate hydrophobicity scale. ***a***, The horizontal levels are the eigenvalues characterizing the intrinsic interactions of the Miyazawa-Jernigan (MJ) matrix (1996 version)[16] of contact interaction energies, in units of gas constant times absolute temperature, RT. The large magnitude of the lowest eigenvalue at -69.7 (corresponding to strongly favorable interactions) dominates significantly over the other 19 eigenvalues, which are all relatively close to zero (dotted line), ranging from -1.9 to +3.4. The lesser eigenvalues account for more specific interactions between residues[20]. It follows that the strengths of pairwise interactions between any two amino acid residues are determined mainly by their projections along the dominant eigenvector, that of the dominant eigenvalue. Such a projection's magnitude quantifies the degree to which a given amino acid residue is represented by the dominant eigenvector. It may be viewed as a crude measure of how much a residue's interactions can be characterized as 'hydrophobic'. ***b***, The projections (vertical lines) along the dominant eigenvector of the 20 amino acid residues (given in one-letter codes) follow an approximate hydrophobicity scale. Residues that are more hydrophobic tend to have larger projections. Wang and Wang's optimized five-letter reduced alphabet (IAGEK) follows a similar pattern — different groupings cover successive ranges of projections, with almost no overlap. (From left to right, the brackets indicate the groups represented by K, E, K, G, A, and I.) The only exception is group 'E', which includes residues D and E, both negatively charged at neutral pH; group E is inserted into the middle of the 'K' group along the projection scale. The mismatch minimization procedure is apparently sufficiently astute, in taking account of less dominant interactions represented by other eigenvalues, so as not to group D and E with positively charged residues such as K and R.

been popular among protein researchers since Miyazawa and Jernigan's[18] detailed analysis in 1985. Godzik *et al.*[19] observed that the MJ contact energies can be deduced largely from the hydrophobicities of the pair of residues forming the contacts. This was subsequently confirmed and further elucidated by a matrix analysis of Li *et al.*[20]. Wang and Wang's reduction results can be understood in a similar light. The MJ matrix has only one dominant eigenvalue (Fig. 2*a*), corresponding to a strongly favorable interaction, as has been discussed[21]. The amino acid components along the dominant eigenvector correspond roughly to a hydrophobicity scale. Interestingly, the grouping in Wang and Wang's optimized five-letter alphabet can be viewed essentially as a classification of amino acids into different ranges of hydrophobicities along this scale (Fig. 2*b*), which is consistent with maximization of interaction heterogeneity in the reduced alphabet since it does not put residues with very different hydrophobicities into the same group (Fig. 1). Therefore, this correspondence is a good indication that their reduction algorithm is handling its task properly.

Whereas Wang and Wang's algorithm is successful in capturing the essentials of the MJ matrix, statistical potentials, such as the MJ potential, have limitations[22]. One problem is that it is difficult to accurately determine repulsive interactions from a native structure database[23], but specific repulsive interactions have been shown to be crucial in simulations of protein folding kinetics[24–26]. At a more fundamental level, residue–residue pairwise interactions by themselves may be inadequate to describe energetics of real proteins[27–29]. Therefore, it remains to be tested whether Wang and Wang's method, as currently applied to pairwise statistical potentials alone, could be successful in predicting experimentally viable reduced alphabets. Nonetheless, it is likely that their mismatch minimization principle, which effectively maximizes interaction heterogeneity, may be generalized to other interaction schemes and thus will be helpful for future experimental designs. Ultimately, however, only new experiments will show if it is possible to construct truly protein-like sequences with an alphabet smaller than seven.

Acknowledgments

I am very grateful to A. R. Davidson for communicating unpublished results and for critically reading this manuscript. I also thank M. Gross, S.L. LaPorte, K.W. Plaxco,D. Thirumalai, and T. Yomo for helpful discussions.

Hue Sun Chan is in the Department of Biochemistry, and Department of Medical Genetics and Microbiology, Faculty of Medicine, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada. email: chan@arrhenius.med.toronto.edu

1. Wong, J.T.-F. *Proc. Natl. Acad. Sci. USA* **72**, 1909–1912 (1975).
2. Wolynes, P.G. *Nature Struct. Biol.* **4**, 871–874 (1997).
3. Meewes, M., Rička, J., de Silva, R., Nyffenegger, R. & Binkert, T. *Macromolecules* **24**, 5811–5816 (1991).
4. Wang, J. & Wang, W. *Nature Struct. Biol.* **6**, 1033–1038 (1999).
5. Morita, K., Simons, E.R. & Blout, E.R. *Biopolymers* **5**, 259–271 (1967).
6. Rao, S.P., Carlstrom, D.E. & Miller, W.G. *Biochemistry* **13**, 943–952 (1974).
7. Munson, M., O'Brien, R., Sturtevant, J.M. & Regan, L. *Protein Sci.* **3**, 2015–2022 (1994).
8. Rojas, N.R.L. *et al. Protein Sci.* **6**, 2512–2524 (1997).
9. Riddle, D.S. *et al. Nature Struct. Biol.* **4**, 805–809 (1997).
10. Brown, B.M. & Sauer, R.T. *Proc. Natl. Acad. Sci. USA* **96**, 1983–1988 (1999).
11. Plaxco, K.W., Riddle, D.S., Grantcharova, V. & Baker, D. *Curr. Opin. Struct. Biol.* **8**, 80–85 (1998).
12. Davidson, A.R., Lumb, K.J. & Sauer, R.T. *Nature Struct. Biol.* **2**, 856–864 (1995).
13. Yamauchi, A. *et al. FEBS Lett.* **421**, 147–151 (1998).
14. Sicheri, F. & Yang, D.S. *Nature* **375**, 427–431 (1995).
15. Schafmeister, C.E., LaPorte, S.L., Miercke, L.J.W. & Stroud, R.M. *Nature Struct. Biol.* **4**, 1039–1042 (1997).
16. Miyazawa, S. & Jernigan, R.L. *J. Mol. Biol.* **256**, 623–644 (1996).
17. Tanaka, S. & Scheraga, H.A. *Macromolecules* **9**, 954–950 (1976).
18. Miyazawa, S. & Jernigan, R.L. *Macromolecules* **18**, 534–552 (1985).
19. Godzik, A., Koliński, A. & Skolnick, J. *Protein Sci.* **4**, 2107–2117 (1995).
20. Li, H., Tang, C. & Wingreen, N.S. *Phys. Rev. Lett.* **79**, 765–768 (1997).
21. Thirumalai, D. & Klimov, D.K. *Fold. Des.* **3**, R112–R118 (1998).
22. Thomas, P.D. & Dill, K.A. *J. Mol. Biol.* **257**, 457–469 (1996).
23. Mirny, L.A. & Shakhnovich, E.I. *J. Mol. Biol.* **264**, 1164–1179 (1996).
24. Socci, N.D. & Onuchic, J.N. *J. Chem. Phys.* **103**, 4732–4744 (1995).
25. Chan, H.S. & Dill, K.A. *Proteins Struct. Funct. Genet.* **30**, 2–33 (1998).
26. Sorenson, J.M. & Head-Gordon, T. *Fold. Des.* **3**, 523–534 (1998).
27. Park, B.H., Huang, E.S. & Levitt, M. *J. Mol. Biol.* **266**, 831–846 (1997).
28. Domany, E., Najmanovich, R. & Vendruscolo, M. In *Monte carlo approach to biopolymers and protein folding* (eds Grassberger, P., Barkema, G.T. & Nadler, W.) 194–210 (World Scientific, Singapore; 1998).
29. Betancourt, M.R. & Thirumalai, D. *Protein Sci.* **8**, 361–369 (1999).