Nonproteinogenic deep mutational scanning of linear and cyclic peptides

Joseph M. Rogers*a, Toby Passioura*b, and Hiroaki Sugaab,1

*Department of Chemistry, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan; and bCore Research for Evolutionary Science and Technology, Japan Science and Technology Agency, Saitama 332-0012, Japan

Published online October 9, 2018.

High-resolution structure–activity analysis of polypeptides requires amino acid structures that are not present in the universal genetic code. Examination of peptide and protein interactions with this resolution has been limited by the need to individually synthesize and test peptides containing nonproteinogenic amino acids. We describe a method to scan entire peptide sequences with multiple nonproteinogenic amino acids and, in parallel, determine the thermodynamics of binding to a partner protein. By coupling genetic code reprogramming to deep mutational scanning, any number of amino acids can be exhaustively substituted into peptides, and single experiments can return all free energy changes of binding. We validate this approach by scanning two model protein-binding peptides with 21 diverse nonproteinogenic amino acids. Dense structure–activity maps were produced at the resolution of single aliphatic atom insertions and deletions. This permits rapid interrogation of interaction interfaces, as well as optimization of affinity, fine-tuning of physical properties, and systematic assessment of nonproteinogenic amino acids in binding and folding.

Significance

The 20 proteinogenic amino acids have physicochemical properties that allow peptides and proteins to fold and bind. However, there are numerous unnatural, nonproteinogenic amino acids that may be equally good, or even better, at folding and binding. Exploration of these alternative peptide building blocks has been limited by slow, one-at-a-time synthesis and testing. We describe how, in a single experiment, multiple nonproteinogenic amino acids can be trialed at all positions in a peptide sequence, with thousands of modifications tested in parallel. This permits detailed analysis of how chemical structure relates to function and allows for systematic comparisons of proteinogenic and nonproteinogenic chemistry. Such analysis can guide the improvement of drug-candidate peptides, including the therapeutically promising class of cyclic peptides.

Author contributions: J.M.R. designed research; J.M.R. and T.P. performed research; J.M.R. and T.P. contributed new reagents/analytic tools; J.M.R. analyzed data; and J.M.R., T.P., and H.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1To whom correspondence should be addressed. Email: hsuga@chem.s.u-tokyo.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1809901115/-/DCSupplemental.

Published online October 9, 2018.
Nonproteinogenic deep mutational scanning. (A) Essentially any nonproteinogenic amino acid [activated by cyanomethyl ester (CME) or dinitrobenzyl ester (DBE)] can be loaded onto tRNA by flexizymes and delivered to the ribosome for use during in vitro translation. Translation of a site-saturation mutagenesis mRNA library using genetic code reprogramming allows for a nonproteinogenic amino acid to be incorporated at all positions in a given peptide sequence, and mRNA display can link each mutant peptide to its encoding mRNA. This can be repeated for n nonproteinogenic amino acids (n = 5 shown). Initiating reverse transcription with a barcoded DNA primer allows for the resulting peptide-cDNA products to be pooled and nonproteinogenic mutants deconvoluted after DNA sequencing. (B) The large diverse library from A, containing proteinogenic and nonproteinogenic mutants, can be sorted for binding to a partner protein. Next-generation sequencing of cDNA before (input) and after (output) binding is then performed. For each mutant, i, the fraction of DNA reads, \( f_i \), can be used to calculate an enrichment score, e. (C) Enrichment scores relative to wild-type, E, provide a map of beneficial and deleterious changes to the peptide chemical structure.

Fig. 1. Nonproteinogenic deep mutational scanning. (A) Essentially any nonproteinogenic amino acid [activated by cyanomethyl ester (CME) or dinitrobenzyl ester (DBE)] can be loaded onto tRNA by flexizymes and delivered to the ribosome for use during in vitro translation. Translation of a site-saturation mutagenesis mRNA library using genetic code reprogramming allows for a nonproteinogenic amino acid to be incorporated at all positions in a given peptide sequence, and mRNA display can link each mutant peptide to its encoding mRNA. This can be repeated for n nonproteinogenic amino acids (n = 5 shown). Initiating reverse transcription with a barcoded DNA primer allows for the resulting peptide-cDNA products to be pooled and nonproteinogenic mutants deconvoluted after DNA sequencing. (B) The large diverse library from A, containing proteinogenic and nonproteinogenic mutants, can be sorted for binding to a partner protein. Next-generation sequencing of cDNA before (input) and after (output) binding is then performed. For each mutant, i, the fraction of DNA reads, \( f_i \), can be used to calculate an enrichment score, e. (C) Enrichment scores relative to wild-type, E, provide a map of beneficial and deleterious changes to the peptide chemical structure.

Results

Selection of Nonproteinogenic Amino Acids for Mutagenesis. Hundreds of nonproteinogenic amino acids are accepted by flexizymes and the ribosome (23); a sample set was chosen for nonproteinogenic mutagenesis, focusing on simple, largely nonpolar structures absent from the universal genetic code. This set of 21 amino acids included both alternative side chains and backbone modifications such as N-methyl substitution, disubstitution, and \( \alpha \)-stereochemistry (Fig. 2A). We chose initially to test these in a natural protein-protein interaction that has already been studied using traditional mutagenesis: the interaction between the apoptosis regulatory BH3 domain of PUMA and the folded protein MCL1 (27, 28). BH3 domains are relatively short and unstructured in isolation but will fold to a single \( \alpha \)-helical upon binding (Fig. 2B). This simple structure has made PUMA and its homologs model systems for protein folding upon binding (29, 30) and for exploring the potential of nonproteinogenic amino acids in druglike peptides (16, 31).

Nonproteinogenic Deep Mutational Scanning. The flexizyme protocol allows for facile reprogramming of genetic codes (Fig. 1A). We assembled in vitro translation systems in which methionine was replaced with one of the 21 nonproteinogenic amino acids (Fig. 2A), with high-fidelity incorporation at AUG codons (SI Appendix, Fig. S2). A site-saturation mutagenesis mRNA library for PUMA was added to each of these systems and translated into PUMA peptides. Similar to previous studies (29, 30), a pseudo wild-type sequence was used containing the mutation M144A (henceforth referred to as the wild-type); this mutation prevents oligomerization of PUMA peptides at high (micromolar) concentrations. Each peptide was covalently attached to its encoding mRNA via a puromycin linker, and the pool of peptide-mRNA fusions was reverse transcribed into noncovalent cDNA complexes using barcoded DNA primers (Fig. 1A) (32). Barcoding, which encoded the reprogrammed genetic codes themselves, permitted an outsized 41 amino acid alphabet.

The resulting diverse PUMA peptide library was incubated with immobilized MCL1 and washed, and the bound fraction was recovered (Fig. 1B). Populations before and after binding were enumerated by deep sequencing, and enrichment scores for binding (E) were calculated for every mutant (2) (Fig. 1C). To validate this approach, previously reported \( K_D/\Delta \Delta G \) values (30) were compared with E scores, which revealed that E is a smooth function of MCL1 binding affinity (Fig. 2C). To validate that E is a function of binding for nonproteinogenic mutants, we chemically synthesized additional PUMA mutants and measured binding to MCL1 (SI Appendix, Figs. S3 and S4 and Table S1). For synthetic convenience, the PUMA peptides in this collection were 27 amino acids in length, shorter than the 35-aa peptides used in previous studies (29, 30), and \( \Delta \Delta G \) values were calculated relative to the binding of an equivalent 27-aa wild-type. The link between \( \Delta \Delta G \) and E is maintained for this collection and, importantly, the proteinogenic and nonproteinogenic mutant data overlay (Fig. 2C).

E scores alone are sufficient for analysis of deep mutational scanning data (2–9). However, to aid interpretation of the mutant data and any structure–activity relationships that follow, we chose to calibrate \( \log E \) against a handful of experimentally validated mutants and calculate true thermodynamic parameters, \( \Delta \Delta G/K_D \), for each mutant. Rather than make assumptions about the function linking \( \log E \) and \( K_D \), we chose to use an empirical fit of E and \( \Delta \Delta G \) to calibrate the deep mutational scanning data. The predictive power of our approach was assessed using leave-one-out cross-validation (Fig. 2D).

An exhaustive map of PUMA mutant \( \Delta \Delta G \) was constructed, including all mutations to all 41 proteinogenic and nonproteinogenic amino acids (Fig. 2E). As expected for \( \alpha \)-helical folding, mutations to proline or backbone N-methyl amino acids were highly destabilizing (29) (Fig. 2E and SI Appendix, Fig. S5). As expected for a BH3 domain (27, 28), mutations to the highly conserved D146 were not tolerated (Fig. 2D and SI Appendix, Fig. S6). A145S, which brings PUMA closer to the BH3 consensus of LXXXGD (28), was highly stabilizing. Interestingly, a nonproteinogenic amino acid substitution was the most favorable at many positions in the PUMA sequence (SI Appendix, Fig. S6).

Nonproteinogenic Deep Mutational Scanning of the de Novo Macrocyclic Peptide CP2. In principal, this scanning approach can be applied to sequences that already contain nonproteinogenic amino acids, such as the small de novo macrocyclic peptides discovered using the RaPID system (23–26). We reasoned that saturation mutagenesis could help understand the molecular details behind the potent activities of these macrocycles and might suggest modifications to amino acids known to improve protease resistance or membrane permeability (13). We chose to investigate the peptide CP2 (Fig. 2F), a potent and isoform-selective inhibitor of the KDM4A histone demethylase, which contains a nonproteinogenic \( \delta \)-tyrosine and is cyclized via a nonreducible thioether bond. CP2 binds KDM4A with a 30 nM \( K_D \), forms a small \( \beta \)-sheet-like secondary structure when bound (Fig. 2F and SI Appendix, Fig. S7), and exhibits inhibitory activity with \( \sim 40 \) nM \( IC_{50} \) (26).

Nonproteinogenic deep mutational scanning was performed for CP2. The raw E scores for binding KDM4A correlated well with previously reported inhibitory \( IC_{50} \) values (SI Appendix, Fig. S7). A small collection of CP2 mutants was synthesized, and
measured ΔΔG values for binding (SI Appendix, Fig. S8 and Table S2) were used to calibrate E and calculate ΔΔG for all proteinogenic and nonproteinogenic CP2 mutants (Fig. 2G and SI Appendix, Fig. S7). The map of ΔΔG shows that a critical RSG motif was intolerant to mutation (Fig. 2G and SI Appendix, Fig. S9) and confirms many of the observations from previous structure–activity studies (33). The RSG motif forms the turn of the β-sheet and is deeply buried in KDM4A (Fig. 2F and SI Appendix, Fig. S7).

**Structure–Activity Relationships at Single-Atom Resolution.** The expanded range of amino acid structures and the systematic nature of this mutagenesis allowed for the extraction of detailed structure–activity relationships at every position in each peptide sequence. As an example, progressive deletion of aliphatic carbon atoms from the side chain L141 of PUMA increasingly destabilizes the interaction with MCL1 (Fig. 3), whereas some atom insertion was stabilizing (e.g., the mutant L141Cpa; Fig. 3) (31). At position A144, where there is a nonnatural feature in the sequence of PUMA (M144A, see above), mutants with longer aliphatic side chains were increasingly stabilizing (Fig. 3). Thus, deep mutational scanning was able to identify the vacant hydrophobic pocket left by the M144A mutation. In the cyclic peptide CP2, R10 could be truncated to aliphatic side chains such as Nva without affecting affinity for KDM4A (Fig. 3), suggesting that it is the hydrophobic chain of the arginine that interacts favorably with the partner protein, not the charged head group. Strikingly, at G8, addition of an (L-) side chain (Fig. 3) is not tolerated, but d-alanine is accepted (Fig. 2F and SI Appendix, Figs. S9 and S10). This led us to inspect the phi and psi angles of bound CP2, which identified G8 as being in a region disallowed for L- but permissible for D- amino acids (SI Appendix, Fig. S11).

**Deep Mutational Scanning-Guided Redesign for Improved Affinity.** The information from these nonproteinogenic mutant scans can be used to engineer peptides for increased affinity, assuming...
that the effects of multiple mutants are, to some extent, additive. For example, five individual affinity-enhancing mutants of PUMA were combined into one redesigned peptide, “rPUMA” (Fig. 4A). rPUMA bound MCL1 with improved affinity ($K_D < 0.16$ vs. 4 nM for an equivalent 27-aa wild-type peptide), stronger than any of the individual mutants, and with markedly slower dissociation compared with the unmodified peptide (Fig. 4A).

Deep Mutational Scanning-Guided Tuning of Peptide Physical Properties. Mutations to nonproteinogenic amino acids can engender peptides with beneficial physical properties that can increase in vivo potency (13), provided they do not disrupt the main function of the molecule. For CP2, this has been attempted before using structure-based design (26) (SI Appendix, Fig. S12), whereby a number of nonproteinogenic mutants were chosen to increase protease resistance and/or cellular activity without impacting the bound state, but modest reductions in affinity for KDM4A were observed (26). The deep mutational scanning data corroborated the structure-based design but also highlighted additional beneficial mutations (SI Appendix, Fig. S12): mutants that reduced steric bulk, added $N$-methylation, removed charges or hydrogen bonding groups, and did not affect KDM4A binding. These mutations were combined to produce “rCP2” (Fig. 4B) and remained energetically neutral in combination: rCP2 bound KDM4A with equivalent affinity to CP2 (Fig. 4B) ($K_D = 7.0$ vs. 6.6 nM for the wild-type), in contrast to the modified CP2 peptides generated through structure-based design.

Analysis of Nonproteinogenic Amino Acids in Binding and Folding. The systematic nature of deep mutational scanning has allowed for global analysis of proteinogenic amino acid substitutions (34) (i.e., comparing the amino acids themselves). Here, we could extend this analysis to include nonproteinogenic amino acids. As an example, scanning the α-helical PUMA with mutations to the nonproteinogenic McB showed a pattern of $\Delta \Delta G$ almost identical to the scan with proteinogenic Pro (Fig. 5). As McB is a noncyclic homolog of Pro, this emphasizes that the $N$-substitution, rather than cyclic structure, is the root cause of Pro destabilization of α-helices. Interestingly, certain nonproteinogenic amino acids closely mimic the behavior of a proteinogenic amino. At all positions in the PUMA and CP2 peptides, mutations to $3T_h$ are energetically equivalent to mutations to the structurally similar Phe; likewise for tBu and Leu (Fig. 5 and SI Appendix, Fig. S13).

Discussion By coupling the synthetic abilities of genetic code reprogramming to the massively parallel analysis of deep mutational scanning, we describe a method to exhaustively trial multiple, diverse nonproteinogenic amino acids in the thermodynamics of protein–protein interactions. This approach permits the study of large numbers of nonproteinogenic amino acid mutants in a single experiment, on a scale that would not be possible using classical cycles of synthesis and biophysical analysis.
Certain nonproteinogenic amino acids closely mimic those in the universal genetic code. Graph shows ΔΔG mutant scans for proteinogenic (black) and nonproteinogenic amino acids (rainbow). MeB shows a similar ΔΔG profile to Pro in the folding and binding of α-helical PUMA (Top), whereas the ΔΔG profile of 3Th mimics Phe (Middle) and tBu mimics Leu (Bottom), in both PUMA and β-sheet CP2. Errors in ΔΔG propagated from SDs in log₂E.
Here, we examined 1,360 mutants of PUMA (714 nonproteinogenic) and 468 mutants of CP2 (240 nonproteinogenic). This is not taken advantage of the maximum mRNA display library size (>10^10) or the full throughput of next-generation DNA sequencing. Therefore, nonproteinogenic deep mutational scanning can be expanded to study multiple mutants or longer protein sequences. This method can be applied to other natural, de novo discovered or designed peptides or proteins that function through binding or through binding and folding.

The ability to rapidly generate high-resolution structure–activity maps will be particularly useful in the development of drug-candidate peptides, as these information-rich datasets can be used to guide modifications for greater binding and potency. In essence, this approach is a highly parallelized version of hit-to-lead exploration in small molecule drug development. Lastly, this method can be extended to include any number of nonproteinogenic amino acids tolerated by the ribosome, a set that contains hundreds of structures and continues to expand (23, 35, 36). Deep mutational scanning can be used to comprehensively assess these nonproteinogenic amino acids for use in binding, folding, and protein engineering.

Materials and Methods

Nonproteinogenic amino acids for mutagenesis (ester activated) were loaded onto tRNA<sup>Met</sup> by using flexizymes (21) (SI Appendix, Table S3). Nonproteinogenic amino acids for translation initiation (N-acetyl-<i>L</i>-phenylalanine for PUMA and N-chloroacetyl-<i>L</i>-tyrosine for CP2) were loaded onto tRNA<sup>Met</sup> by using the flexizyme eFx (22). Loaded tRNA were separately added to methionine-deficient in vitro translation systems, along with template-saturation mRNA libraries (with C-terminal HA tag) for PUMA or CP2, and translated into peptides according to an mRNA display protocol (25). Reverse transcription was carried out using barcoded DNA primers unique for each reprogrammed genetic code, and the products from each genetic code were pooled. Nonproteinogenic amino acids were incorporated with high fidelity (SI Appendix, Fig. S2), but to account for any differences in translation efficiencies, cDNA-linked peptide mutants (with C-terminal HA tag) were purifying using anti-HA magnetic beads to remove any incompletely translated by-products.

Twenty microliters of an approximately 200 nM library of anti-<i>HA</i>-purified cDNA-linked peptide mutants (PUMA or CP2) was incubated with 200 nM protein binding partner biotin-streptavidin immobilized onto magnetic beads (MCL1 or KDM4A). A 200 nM concentration of partner protein was low enough to produce a broad distribution of log<sub>E</sub>, while high enough to avoid excessive PCR cycles during DNA recovery (SI Appendix, Fig. S14). Binding was allowed to reach equilibrium (3 h at 25 °C), and the beads were twice washed with buffer (20 μL each), allowing binding to reach equilibriuim after each buffer addition (three 3-h incubations in total). Samples of the anti-<i>HA</i>-purified cDNA libraries before and after binding were prepared for Illumina sequencing. DNA reads were analyzed using a modified version of the Enrich pipeline (2), calculating enrichment scores (E) for each proteinogenic and nonproteinogenic mutant. Repeat experiments of the incubation, washing, and PCR amplification, (starting from the same library preparation, n = 3 for PUMA, n = 4 for CP2) were averaged to give the reported log<sub>E</sub> values, and the SDs were used to estimate the error in log<sub>E</sub>. Depending on the function ΔΔ<sub>G</sub> = f(log<sub>E</sub>), errors were propagated appropriately to give reported errors in ΔΔ<sub>G</sub>.

A selection of proteinogenic and nonproteinogenic mutants of PUMA and CP2, covering the range of E scores, was synthesized using solid-phase peptide synthesis, and binding to MCL1 and KDM4A, respectively, was tested using surface plasmon resonance. The resulting K<sub>E</sub>/ΔΔ<sub>G</sub> values, plus any published values (30), were used to find empirical relationships between <i>E</i> and ΔΔ<sub>G</sub>, and these functions were used to calculate K<sub>P</sub>/ΔΔ<sub>G</sub> for all PUMA and CP2 mutants.

Acknowledgments

We thank A. Kawamura for the gift of biotinylated KDM4A. This study was supported by Grants-in-Aid for Japan Society for the Promotion of Science Fellows (P13766) (to J.M.R.), a Joint French National Research Agency/Japan Science and Technology Agency Grant (ANR-14-JTC-2014-003 and JST-SICORP), a Japan Science and Technology Agency Core Research for Evolutionary Science and Technology Grant (JPMJCR12L2), and a Japan Agency for Medical Research and Development Grant (JP18am0301001).

1. Weiss GA, Watanabe CK, Zhong A, Goddard A, Sidhu SS (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. Proc Natl Acad Sci USA 97:8950-8954.
2. Fowler DM, et al. (2010) High-resolution mapping of protein sequence-function relationships. Nat Methods 7:741-746.
3. Rocklin GJ, et al. (2017) Global analysis of protein folding using massively parallel functional scanning. Nat Methods 14:479-484.
4. Araya GL, et al. (2012) A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc Natl Acad Sci USA 109:16858-16863.
5. Olson CA, Wu NC, Sun R (2014) A comprehensive biophysical description of pairwise thermodynamic relationships. Nat Methods 11:1238-1244.
6. van Rosmalen M, et al. (2017) Affinity maturation of a cyclic peptide handle for high-affinity drug delivery. ChemBioChem 18:2130-2137.
7. Whitehead TA, et al. (2012) Optimization of affinity, specificity and function of de novo macrocyclic peptide inhibitors against a ubiquitin ligase uncovered from a ribosome-expressed de novo library. Chem Biol 18:1562-1570.
8. Kawamura A, et al. (2017) Highly selective inhibition of histone Demethylases by de novo macrocyclic peptide inhibitors. Nat Commun 8:14773.
9. Rezaei Araghi R, Ryan JA, Leal A, Keating AE (2016) Rapid optimization of McI inhibitors using stapled peptide libraries including non-natural side chains. ACS Chem Biol 11:1238-1244.
10. Tjhung KF, et al. (2016) Silent encoding of chemical post-translational modifications in phage-displayed libraries. J Am Chem Soc 138:12428-12429.
11. Tjhung KF, et al. (2017) Stepwise evolution improves identification of diverse active Peptide libraries. Proc Natl Acad Sci USA 114:14136-14141.
12. Checco JW, et al. (2015) Beyond cyclosporine A: Conformation-dependent passive therapeutic antibodies using deep mutational scanning. J Mol Evol 78:64-77.
13. Murakami H, Ohta A, Ashigai H, Suga H (2006) A highly flexible tRNA acylation mechanism after each buffer addition (three 3-h incubations in total). Samples of the anti-<i>HA</i>-purified cDNA libraries before and after binding were prepared for Illumina sequencing. DNA reads were analyzed using a modified version of the Enrich pipeline (2), calculating enrichment scores (E) for each proteinogenic and nonproteinogenic mutant. Repeat experiments of the incubation, washing, and PCR amplification, (starting from the same library preparation, n = 3 for PUMA, n = 4 for CP2) were averaged to give the reported log<sub>E</sub> values, and the SDs were used to estimate the error in log<sub>E</sub>. Depending on the function ΔΔ<sub>G</sub> = f(log<sub>E</sub>), errors were propagated appropriately to give reported errors in ΔΔ<sub>G</sub>.

A collection of proteinogenic and nonproteinogenic mutants of PUMA and CP2, covering the range of E scores, was synthesized using solid-phase peptide synthesis, and binding to MCL1 and KDM4A, respectively, was tested using surface plasmon resonance. The resulting K<sub>E</sub>/ΔΔ<sub>G</sub> values, plus any published values (30), were used to find empirical relationships between <i>E</i> and ΔΔ<sub>G</sub>, and these functions were used to calculate K<sub>P</sub>/ΔΔ<sub>G</sub> for all PUMA and CP2 mutants.