PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SHORT COMMUNICATION

ScaffoldSeq: Software for characterization of directed evolution populations

Daniel R. Woldring, Patrick V. Holec, and Benjamin J. Hackel*

Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455

ABSTRACT

ScaffoldSeq is software designed for the numerous applications—including directed evolution analysis—in which a user generates a population of DNA sequences encoding for partially diverse proteins with related functions and would like to characterize the single site and pairwise amino acid frequencies across the population. A common scenario for enzyme maturation, antibody screening, and alternative scaffold engineering involves naïve and evolved populations that contain diversified regions, varying in both sequence and length, within a conserved framework. Analyzing the diversified regions of such populations is facilitated by high-throughput sequencing platforms; however, length variability within these regions (e.g., antibody CDRs) encumbers the alignment process. To overcome this challenge, the ScaffoldSeq algorithm takes advantage of conserved framework sequences to quickly identify diverse regions. Beyond this, unintended biases in sequence frequency are generated throughout the experimental workflow required to evolve and isolate clones of interest prior to DNA sequencing. ScaffoldSeq software uniquely handles this issue by providing tools to quantify and remove background sequences, cluster similar protein families, and dampen the impact of dominant clones. The software produces graphical and tabular summaries for each region of interest, allowing users to evaluate diversity in a site-specific manner as well as identify epistatic pairwise interactions. The code and detailed information are freely available at http://research.cems.umn.edu/hackel.

Proteins 2016; 84:869–874.

Key words: bioinformatics; high-throughput; sequence analysis; family clustering; epistasis; software.

INTRODUCTION

Sequence analysis of diverse protein populations with related functions is valuable in characterizing antibody[1–4] repertoires, evaluating homologs (such as for consensus design[5]), guiding combinatorial library design for *de novo* protein discovery,[6] and performing deep mutational scanning[7,8] to elucidate evolution, for example, of stability, binding, or catalysis. Sequence analysis can identify—on a sitewise or multi-site motif level—amino acid frequencies that are consistent with the discovery and evolution of stable, functional molecules. These amino acid frequencies can be implemented combinatorially in libraries or precisely in clones. The increased availability of broad data sets of functionally homologous, partially diverse proteins mirrors the growth in deep sequencing and bioinformatics mining.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Department of Defense; grant number: W81XWH-13-1-0471; Grant sponsor: National Institutes of Health; grant number: R21 EB019518.

*Correspondence to: Benjamin J. Hackel, Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455. E-mail: hackel@umn.edu

Received 4 November 2015; Revised 8 March 2016; Accepted 18 March 2016
Published online 28 March 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.25040

Realizing benefit from these advances requires techniques and software for efficient, accurate, consistent analysis throughout and across fields.

Here we discuss software to analyze diverse protein populations for such purposes. As a particular type of example, we highlight the analysis of populations of small protein scaffolds (fibronectin type III[6] and Gp2 domains[9]) in which three or two regions, respectively, were highly diversified and evolved for various binding functions. The software input is DNA sequences (FASTA/FASTQ); for example, a population encoding for protein domains engineered to bind various epitopes on an antigen. The primary outputs are sitewise amino acid frequency matrices, pairwise epistasis analyses—including epistatic frequency distributions and identification of key positive and negative correlates—and metrics of sequence diversity. The analysis workflow differentiates itself from existing tools and methods of others[10–17] by being customizable, via a dynamic, easily-navigated user interface, and allows removal of background sequences (e.g., nonfunctional clones unintentionally isolated during a protein library screen), evaluation and clustering of highly similar sequences, and dominant clone weight dampening. Output files are generated as graphical summaries and in comma-separated value format to facilitate downstream application and project-specific data interrogation (Fig. 1). Along with the annotated script, an accompanying *Software Walkthrough* provides a detailed guide for users as exemplified by Gp2-directed evolution analysis. Beyond the ligand-specific scientific value of sitewise and pairwise amino acid frequency data, analysis of the affibody,[18] fibronectin, and Gp2 ligand evolution data reveals the benefits of variable dampening of dominant clones.

MATERIALS AND METHODS

Interface design

Scripts, developed using Python (v2.7) with default libraries to ease portability, are compatible with Windows 7/8, Mac OS X, and Linux OS. An intuitive interface guides the user through the sequence analysis menus and allows for command line execution. While workflow settings are customizable for essentially any protein, default profiles for fibronectin,[6] affibody,[18] DARPin,[19] knottin (kalata B1),[20] and Gp2[9] are included. Each unique profile containing scaffold-specific parameters and settings declared within the job submission menu are saved via Python's *pickle* module to separately store setting arrays outside the main interface. This allows users to easily retrieve and view settings from prior analyses. A step-by-step tutorial is included as a resource to guide users through selecting the appropriate job settings (see *Software Walkthrough*).

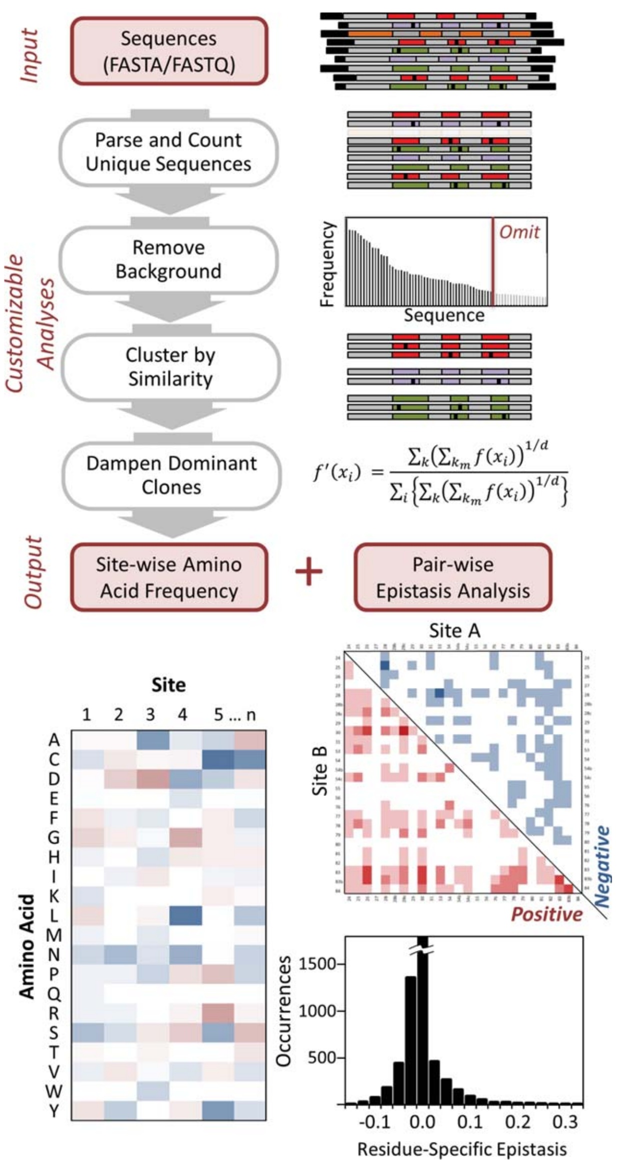
Input
Sequences (FASTA/FASTQ)
Parse and Count Unique Sequences
Customizable Analyses
Remove Background
Frequency
Omit
Sequence
Cluster by Similarity
Dampen Dominant Clones
$f'(x_i) = \frac{\Sigma_k(\Sigma_{k_m} f(x_i))^{1/d}}{\Sigma_i\{\Sigma_k(\Sigma_{k_m} f(x_i))^{1/d}\}}$
Output
Site-wise Amino Acid Frequency
+
Pair-wise Epistasis Analysis
Site A
Site B
Negative
Positive
Site
1 2 3 4 5 … n
Amino Acid
A C D E F G H I K L M N P Q R S T V W Y
Occurrences
1500
1000
500
-0.1 0.0 0.1 0.2 0.3
Residue-Specific Epistasis

Figure 1

ScaffoldSeq evaluates high-throughput sequence data to characterize the diversity within directed evolution and natural populations. The regions of interest within a collection of proteins are identified using tunable similarity thresholds associated with reference sequences. Background sequences are quantified and excluded from the analysis (optional). Highly similar clones are clustered. Population heterogeneities are further elucidated by dampening the impact of highly frequent unique clones. The software generates output files that detail sitewise amino acid distributions and identify pairwise epistasis.

Sequence alignment

Quality, relevant sequences are parsed from the FASTA/FASTQ input file by looping through individual reads, locating conserved anchors at the 5′ and 3′ ends of the gene of interest, and removing segments outside of the anchors. Trimmed sequences of acceptable size (dictated by gene length and allowed length variation, such as from loop length diversity) are aligned to a user-input reference sequence using a cross-correlation test with Python's *difflib* module. This alignment depends on a sufficient fraction of matching nucleotides within the framework regions, directly outside of the diversified regions of interest, having at least 80% matching nucleotides. While the algorithm does not enable FASTQ read quality filtering, the framework alignment threshold can be adjusted by the user with discretion (see *Software Walkthrough*). This has the effect of searching along a trimmed sequence for the transition between a conserved region and a diversified region of interest. The identification method provides rapid location of diversified regions even in cases where the composition and length of the region of interest are not specifically known, which in turn has the advantage of accurately accounting for specific library designs involving loop length diversity within antibody fragments[21,22] and small scaffolds such as fibronectin[23] and Gp2.[9] Length differences are accounted for by inclusion of gap indicators, "-", to maintain alignment in conserved regions.

Background consideration

Background sequences or noise should be accounted for based on the specific experiments that yielded the sequence set. In directed evolution, functional clones are often isolated by survival selections or screening via genotype-phenotype display technologies[24–27] coupled with cell panning,[28] bead sorting,[29] or flow cytometry.[30] Survival and selection techniques yield a small fraction of false positives or background, which can be quantified via control experiments (non-random "false" positives resulting from poor assay design must be accounted for separately). These unwanted, random clones, which are the rarest sequences within the data set, are excluded from the analysis by removing either a user-defined fraction of the rarest sequences (e.g., 2% for example directed evolution population isolated by yeast display and magnetic bead sorting) or all sequences with fewer than a user-defined number of occurrences. Accounting for assay-specific background levels has the additional advantage of sufficiently compensating for read error rates of next-generation sequencing platforms, which tend to be <1%.[31]

Family clustering and frequency calculation with dampening

Sampling bias introduced by dominant motifs and selection methods are detrimental to the statistical analysis of raw sequence data.[32,33] To compensate for these biases, ScaffoldSeq allows for separate correction factors to be associated with motif clustering and individual sequence contributions. Similar sequences (default: 80% identity in diversified sites) can be clustered into families, which facilitate identification of key motifs. Families provide an additional metric for population diversity (sequences, unique sequences, and families), which, in turn, enables broader characterization of the sequence set. Notwithstanding, the diversity of a population can be obscured by a given family similar to how a prominent sequence can upstage all other members of a family. To correct for these imbalances and rather emphasize the heterogeneity of functional clones when appropriate, a dampening exponent can be applied to the frequency of unique sequences to lower the impact of dominant sequences[6,9] [Eq. (1)].

$$f'(x_i) = \frac{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}}{\sum_i \left\{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}\right\}} \qquad (1)$$

where $f(x_i)$ is the frequency of amino acid i at site x; k_m is the mth sequence in family k; and f' is the dampened frequency with dth root dampening.

Traditional sequence analysis often treats each sequence as a distinct solution to a problem. However, within a population, two non-identical, but highly similar sequences may share a common structural or functional motif, akin to providing comparable solutions to the same problem. By lowering the *Sequence Similarity Threshold*, the ScaffoldSeq algorithm defines a broader range of related sequences to be a common solution. The contribution of each common solution (i.e., dominant clones and their common-motif variants) can be tuned to suit the needs of the analysis by using family clustering in combination with dampening.

The *Frequency Dampening Power* ($1/d$) will typically be within the range of 0.25–1. As this value approaches zero, the data set will be treated as though all duplicate sequences were removed. A value of 1 has the effect of weighting all sequences equally and, consequently, negates all impact of clustering, irrespective of the *Sequence Similarity Threshold*. *Frequency Dampening Power* of 0.5 is suggested for sequence data sets that contain a relatively high number of occurrences for a few dominant clones. Sensitivity analyses (Fig. 2) guide selection of appropriate parameter value.

Pairwise interactions

Sitewise amino acid frequencies, $f(x_i)$, are most relevant when each site acts independently. In reality, cohorts of residues are likely to interact under evolutionary pressure.[35–37] Therefore, ScaffoldSeq also compares pairwise residue distributions from full-length evolved sequences relative to those predicted by the region-specific independent frequency matrix, which empowers identification of positive and negative epistasis[38] (Supporting Information Figs. S1 and S2). Specifically, mutual information, MI(x,y), is calculated for each pair of sites, x and y [Eq. (2)]. Each of the 400 possible amino acid

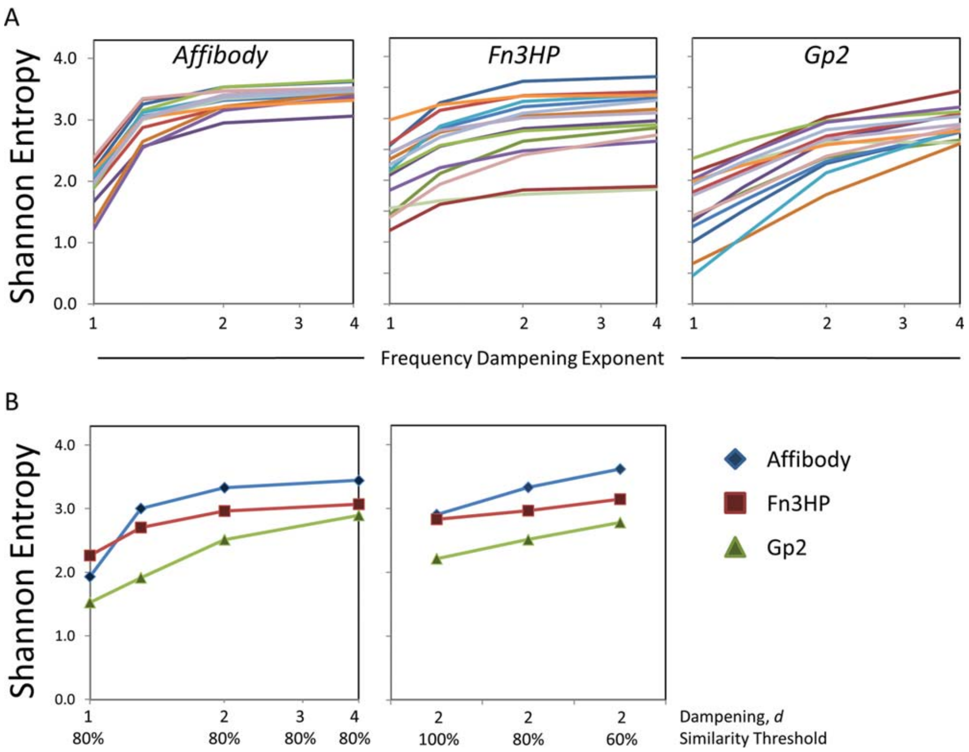

Figure 2

Sensitivity analyses. **A**: Apparent sitewise diversity (i.e., Shannon entropy[34]) within binding populations from recent studies is shown as a function of dampening the effect of dominant sequences by modulating the frequency dampening parameter, d [Eq. (1)]. The family cluster similarity threshold is set at 80% for data from all three scaffolds: hydrophilic fibronectin (Fn3HP),[6] Gp2,[9] and affibody (unpublished). Data analyses were conducted with frequency dampening coefficient ranging from 1 (no dampening) to 4 (heavy dampening). Each line shows the Shannon entropy of a single site of interest. Conclusions: Accounting for dominant clones, via frequency dampening, has a greater impact on less diverse populations. Dampening uniquely affects each site, as demonstrated by changes in the rank order between levels of dampening. **B**: Shannon entropy, averaged across all diversified sites for each scaffold, is shown for a range of frequency dampening coefficients and sequence similarity clustering thresholds. On the left, the similarity threshold is set at 80%. Data analyses were conducted with frequency dampening coefficient ranging from 1 (no dampening) to 4 (heavily dampened). On the right, the frequency dampening coefficient is set to 2 and the similarity threshold is varied. Shannon entropy ($H = -\sum_{i=1}^{20} f_i \log_2 f_i$) , where f_i is the fraction of amino acid i at a particular site) describes relative diversity within the range of 0 (fully conserved) to 4.3 (5% of each amino acid). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

combinations for a site-pair are evaluated based on the predicted sitewise frequency product, $f(x_i)f(y_j)$ and experimentally observed pairwise frequency, $f(x_i,y_j)$ [Supporting Information Eq. (S1)]. For each of the amino acid-specific contributions, positive values indicate the propensity of two mutations to occur more often than would be predicted by a sitewise frequency analysis alone. Mutations that do not occur within the data set are excluded. The summation of these residue-specific values yields the mutual information for that site-pair. The amino acid-specific components of the mutual information calculation are also output to facilitate epistasis analysis. Mutual information from raw sequences is vulnerable to inaccuracies driven by sequence alignments in multiple scenarios. Broadly diverse or quickly evolving sites with high entropy tend to yield larger mutual information scores, irrespective of paired interaction.[39] Countering these effects through normalization techniques has been discussed.[40–43] Bias is also propagated through redundant sequences and prominent families of highly similar clones.[33] Previous corrective efforts include removing all duplicates[11] and weighting sequence counts inversely proportional to the total cluster size.[44] Additionally, high background can arise from small samples sizes,[45] but can be offset by low count correction.[44] While the ScaffoldSeq algorithm largely overcomes these issues via dampening, clustering, and background removal, the mutual information output data incorporates the average product corrected method[43] [Eq. (3)]

$$\mathrm{MI}(x,y)=\sum_i\sum_j f(x_i,y_j)\log_2\frac{f(x_i,y_j)}{f(x_i)f(y_j)} \tag{2}$$

$$\mathrm{MI}_p(x,y)=\mathrm{MI}(x,y)-\frac{\mathrm{MI}(x,*)\mathrm{MI}(*,y)}{\mathrm{MI}(*,*)} \tag{3}$$

where MI(x,*) and MI(*,y) are the average mutual information values of site-pairs involving site x and y,

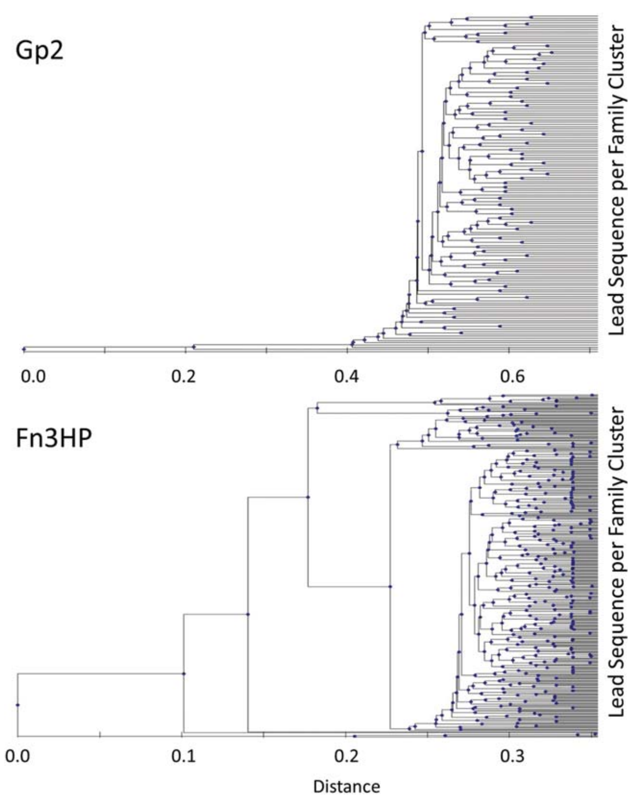

Figure 3

Family clustering performed within ScaffoldSeq can be further evaluated in a phylogenetic tree visualization. The list of dominant sequences within each clustered family, directly output from ScaffoldSeq in the .csv file, were input into the *seqpdist* and *seqlinkage* functions within MATLAB. Horizontal lines on the far right indicate each unique sequence. The *x* axis quantifies the distances between sequences based on the Jukes–Cantor method and blosum50 scoring matrix. The data sets originate from evolved populations of high affinity binders, sequenced by Illumina MiSeq, were analyzed using ScaffoldSeq. Analysis parameters for the Gp2 scaffold (top) and hydrophilic fibronectin (Fn3HP; bottom) populations included a clustering threshold of 0.85 and 0.95, respectively and an assay background filter of 10 for both datasets. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

respectively. MI(*,*) is the average mutual information values across all site-pairs.

RESULTS AND DISCUSSION

ScaffoldSeq has been developed, optimized through extensive testing, and made available for public use along with documentation to facilitate implementation across various applications for the aforementioned functions.[46] Upon completion of the sequence analysis, ScaffoldSeq publishes data in comma-separated value format summarizing each stage of the workflow depicted in Figure 1. Output data include the total number of quality sequences that were parsed by ScaffoldSeq, number of occurrences for each unique protein sequence, background threshold count by which sequence removal was determined, dampening coefficient applied (*d*), number of family clusters, and protein sequence and frequency for unique clones within each cluster. Sitewise amino acid counts and frequency distributions are presented in matrix form. Following this, lead clones from the population (i.e., most prevalent clone from each sequence cluster) are enumerated in rank order. Pairwise similarity distances are computed for each lead clone based on the relative Hamming distance as well as the revised BLOSUM64 score matrix.[47] Many analyses are afforded by the ScaffoldSeq output beyond those included in the default package. One potential application is demonstrated whereby the list of lead clones from each family is used to construct a phylogenetic tree, allowing for visual assessment of high level diversity (Fig. 3). Pairwise diversity analysis, evaluated via mutual information and residue-specific epistasis, is conducted for all 400 pairs of residues, *i* and *j*, at all pairs of sites, *x* and *y* (further discussion in Supporting Information Figs. S1 and S2).

High-throughput evolution (directed or natural) and deep sequencing can substantially advance our knowledge of sequence–function relationships to yield improved mutant or combinatorial library designs. In addition to enlightening analysis of single proteins, sitewise (single and paired) consideration of inter- and intra-molecular interactions—quantified via evolutionary prevalence—can aid combinatorial library designs for *de novo* protein discovery. ScaffoldSeq facilitates such analyses.

REFERENCES

1. Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, Schroeder Jr HW, Kirkham PM. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J Mol Biol 2003;334:733–749.
2. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dörner T, Andrews SF, Wilson PC, Hunicke-Smith SP, Willson CG, Ellington AD, Georgiou G. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol 2013;31:166–169.
3. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, Chrysostomou C, Hunicke-Smith SP, Iverson BL, Tucker PW, Ellington AD, Georgiou G. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol 2010;28:965–969.
4. Tse E, Lobato MN, Forster A, Tanaka T, Chung GTY, Rabbitts TH. Intracellular antibody capture technology: application to selection of intracellular antibodies recognising the BCR-ABL oncogenic proteina. J Mol Biol 2002;317:85–94.
5. Steipe B, Schiller B, Pluckthun A, Steinbacher S. Sequence statistics reliably predict stabilizing mutations in a protein domain. J Mol Biol 1994;240:188–192.
6. Woldring DR, Holec PV, Zhou H, Hackel BJ. High-Throughput ligand discovery reveals a sitewise gradient of diversity in broadly evolved hydrophilic fibronectin domains. PLoS One 2015;10: e0138956.

7. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods 2014;11:801–807.
8. Tripathi A, Varadarajan R. Residue specific contributions to stability and activity inferred from saturation mutagenesis and deep sequencing. Curr Opin Struct Biol 2014;24:63–71.
9. Kruziki MA, Bhatnagar S, Woldring DR, Duong VT, Hackel BJ. A 45-Amino-acid scaffold mined from the PDB for high-affinity ligand engineering. Chem Biol 2015;22:946–956.
10. Fowler DM, Araya CL, Gerard W, Fields S. Enrich: software for analysis of protein function by enrichment and depletion of variants. Bioinformatics 2011;27:3430–3431.
11. Kim T, Tyndel MS, Huang H, Sidhu SS, Bader GD, Gfeller D, Kim PM. MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. Nucleic Acids Res 2012;40:e47.
12. Bloom JD. Software for the analysis and visualization of deep mutational scanning data. BMC Bioinformatics 2015;16:168.
13. Matochko WL, Chu K, Jin B, Lee SW, Whitesides GM, Derda R. Deep sequencing analysis of phage libraries using Illumina platform. Methods 2012;58:47–55.
14. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpää MJ, Bonke M, Palin K, Talukder S, Hughes TR, Luscombe NM, Ukkonen E, Taipale J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res 2010;20:861–873.
15. Alam KK, Chang JL, Burke DH. FASTAptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. Mol Ther Acids 2015; 4:e230.
16. Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, Calloud S, Kosco-Vilbois M, Fischer N. Deep sequencing of phage display libraries to support antibody discovery. Methods 2013;60: 99–110.
17. Dickson RJ, Gloor GB. Bioinformatics identification of coevolving residues. Methods Mol Biol 2014;1123:223–243.
18. Feldwisch J, Tolmachev V, Lendel C, Herne N, Sjöberg A, Larsson B, Rosik D, Lindqvist E, Fant G, Höidén-Guthenberg I, Galli J, Jonasson P, Abrahmsén L. Design of an optimized scaffold for affibody molecules. J Mol Biol 2010;398:232–247.
19. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A. Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. J Mol Biol 2003;332:489–503.
20. Getz JA, Rice JJ, Daugherty PS. Protease-resistant peptide ligands from a knottin scaffold library. ACS Chem Biol 2011;6:837–844.
21. Mahon CM, Lambert MA, Glanville J, Wade JM, Fennell BJ, Krebs MR, Armellino D, Yang S, Liu X, O'Sullivan CM, Autin B, Oficjalska K, Bloom L, Paulsen J, Gill D, Damelin M, Cunningham O, Finlay WJ. Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. J Mol Biol 2013;425:1712–1730.
22. Lamminmäki U, Paupério S, Westerlund-Karlsson A, Karvinen J, Virtanen PL, Lövgren T, Saviranta P. Expanding the conformational diversity by random insertions to CDRH2 results in improved antiestradiol antibodies. J Mol Biol 1999;291:589–602.
23. Hackel BJ, Kapila A, Wittrup KD. Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. J Mol Biol 2008;381:1238–1252.
24. Parmley SF, Smith GP. Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. Gene 1988;73:305–318.
25. Seelig B. mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. Nat Protoc 2011;6:540–552.
26. Mattheakis LC, Bhatt RR, Dower WJ. An in vitro polysome display system for identifying ligands from very large peptide libraries. Proc Natl Acad Sci USA 1994;91:9022–9026.
27. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol 1997;15:553–557.
28. Marks JD, Ouwehand WH, Bye JM, Finnern R, Gorick BD, Voak D, Thorpe SJ, Hughes-Jones NC, Winter G. Human antibody fragments specific for human blood group antigens from a phage display library. Bio/Technology 1993;11:1145–1149.
29. Ackerman M, Levary D, Tobon G, Hackel B, Orcutt KD, Wittrup KD. Highly avid magnetic bead capture: an efficient selection method for de novo protein engineering utilizing yeast surface display. Biotechnol Prog 2009;25:774–783.
30. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD. Isolating and engineering human antibodies using yeast surface display. Nat Protoc 2006;1:755–768.
31. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012;13:1.
32. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 2011;108:E1293–E1301.
33. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
34. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:623–656.
35. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM, Petropoulos CJ, Bonhoeffer S. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nat Genet 2011; 43:487–489.
36. Gong LI, Bloom JD. Epistatically interacting substitutions are enriched during adaptive protein evolution. PLoS Genet 2014;10: e1004328.
37. de Juan D, Pazos F, Valencia A. Emerging methods in protein coevolution. Nat Rev Genet 2013;14:249–261.
38. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics 2005;21:4116–4124.
39. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins Struct Funct Genet 2004;56:211–221.
40. Brown CA, Brown KS. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my!. PLoS One 2010;5:e10779.
41. Durani V, Magliery TJ. Protein engineering and stabilization from sequence statistics: variation and covariation analysis., Methods Enzymol 2013;523:237–256.
42. Little DY, Chen L. Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. Shiu S-H, Ed. PLoS One 2009;4: e4762
43. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–340.
44. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 2009;25:1125–1131.
45. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 2000;17:164–178.
46. Woldring DR, Holec P V, Hackel BJ. Hackel Lab GitHub. April 5, 2016. Available at: https://github.com/HackelLab-UMN.
47. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. BLOSUM62 miscalculations improve search performance. Nat Biotechnol 2008;26:274–275.