Request Information

Permalink

Selection Of DNA-Encoded Libraries For Membrane-Permeable Scaffolds

Tech ID: 34415 / UC Case 2024-825-0

BACKGROUND

Combinatorial encoded library technologies can provide a set of tools for discovering protein-targeting ligands (molecules) and for drug discovery. These techniques can accelerate ligand discovery by leveraging chemical diversity achievable through genetically encoded combinatorial libraries, for example, by combinatorial permutation of chemical building blocks.

Although display technologies such as mRNA and phage display use biological translation machinery to produce peptide-based libraries, hits from these libraries often lack key drug-like properties, for example, cell permeability. This limitation can arise from the peptide backbone's inherent polarity and the tendency to select compounds with polar/charged side chains. Backbone N-methylation can increase scaffold lipophilicity in mRNA display; however, codon table constraints can necessitate longer sequences to fully utilize the available space.

DNA-encoded libraries (DELs) offer an alternative approach towards discovering hits against drug targets. However, like other encoded library techniques, DELs face significant obstacles in affinity selections, which tend to enrich library members bearing polar and/or charged moieties, which can have low (poor) passive cell membrane permeability, especially in larger molecular weight libraries, resulting in hits with poor drug-like properties. This selection bias is especially problematic for larger constructs beyond the rule of 5, where fine-tuning lipophilicity can be critical.

Furthermore, DNA-encoded libraries can be of low quality. Although algorithmic predictions of lipophilicity exist, these two-dimensional (2D) atomistic calculations cannot capture conformational effects exhibited by larger molecules like peptide macrocycles. Despite over a decade of DEL technology development, no method exists to measure physical properties of encoded molecules across an entire DNA-encoded library. That is, successful translation of hits from encoded library selections can be impeded by low quality libraries and enrichment of highly polar members which tend to have poor passive cell permeability, especially for larger molecular weight libraries.

DELs are produced through split-pool synthesis with DNA barcoding to encode the building block of each chemical step. Although this approach can draw on a large number of building blocks and allow for the formation of non-peptidic libraries with a large number of members, synthetic challenges persist. The formation of DELs can be synthetically inefficient. Truncations multiply (are compounded) throughout synthesis, reducing the representation of properly synthesized constructs. Although strategies to improve library purity, to enable reaction monitoring for macrocycle formation, and to identify problematic chemistry affecting DNA tag amplification may be applied, a direct method for assessing DEL quality on a library-wide basis has yet to be developed.

TECHNOLOGY DESCRIPTION

Scott Lokey's lab at UC Santa Cruz first developed a new model involving liquid crystal mass spectrometry and UV detection for the assessment of drug like properties and quality control of a DNA encoded library

CONTACT

Jeff M. Jackson
jjackso6@ucsc.edu
tel: .

INTRODUCING
UC TechAlerts
New technology matches delivered to your email at your preferred schedule
🔍 SEARCH ▸ 🔖 SAVE SEARCH
Learn More ▸

INVENTORS

▸ Lokey, R. Scott

OTHER INFORMATION

KEYWORDS

library, compound library, DNA encoded library, DEL, quality control, LCMS, UV, lipophilicity, prediction of drug-like qualities

CATEGORIZED AS

▸ **Materials & Chemicals**
  ▸ Biological
  ▸ Chemicals
▸ **Medical**
  ▸ New Chemical Entities, Drug Leads
  ▸ Screening

RELATED CASES

2024-825-0

(DEL) and then synthesized a new DEL of cyclic peptides containing diverse backbone elements to test it on. They showed that the model could separate DNA-conjugated macrocycles chromatographically in the context of a fully encoded library as well as measure the lipophilicity of DNA-bound compounds, both in the context of a fully encoded library.

The DNA-encoded library (DEL) was synthesized as a three-cycle DEL of click-cyclized peptides formed of building blocks (BBs) having a variety of lengths and backbone geometries. The DNA encoded library was synthesized with 32 building blocks (BBs) in the first and second cycles (constituent regions, which can result from steps in a chemical synthetic scheme) and 58 building blocks in the third cycle, resulting in a 59,392-member library. The building blocks were selected to prioritize scaffold diversity and span a wide range of lipophilicity.

Building blocks (BBs) included fluorenylmethoxycarbonyl protected (Fmoc) a-amino acids, N-methyl a-amino acids, J3-homoamino acids, N-alkyl glycines (peptoids), and backbone-diverse tripeptides. All stereochemical variants of each amino acid, N-methyl amino acid, J3-homoamino acid, N-alkyl glycine (peptoid), N-functionalized amino acid, and tripeptide were included to maximize the range of backbone geometries within the library and to sample a wide range of backbone conformational space. The aliphatic side chains of the amino acids were selected to feature a variety of branching patterns (degrees of branching), which may produce different effects of hydrophobic shielding dependent on the relative position, and to sample a range of lipophilicities. Incorporation of single amino acids and tripeptides at each position yielded four different ring sizes: 5mers, 7mers, 9mers, and llmers, which span a molecular weight range between 500 and 1200 Da.

Controls were built into the library to facilitate compound sequencing and deconvolute sequence count chromatograms. Each building block (BB) at each position (cycle) was encoded with three unique DNA tags at each cycle (step) of chemistry, resulting in 27 unique (redundant) encoding tag sequences for each compound. Additionally, the compounds of the library were closed with a partially single stranded tag bearing both a priming region and an Nl2 unique molecular identifier (UMI). Combined, the tagging in triplicate and the UMI resulted in (theoretically) approximately 453 million unique sequences per compound.

Another control was the inclusion of three null tag sequences at each cycle (step) of chemistry. Incorporating nulls at each cycle (step) of chemistry engineers encoded truncations into the library, which can provide additional chromatographic data for comparison and aid in product identification. That is, in order to track synthetic truncations, at each chemistry cycle (step) three pools received a set of "null" barcodes but were not coupled to a BB at that step. After the three building block cycles, the pooled library was N-terminally modified (acylated) with azido acetic acid. Half of the library was left linear (uncyclized) while the other half was macrocyclized with a copper mediated click cycloaddition (cyclization). Accounting for the null compound diversity, the overall theoretical diversity for (number of compounds in) either the

linear or cyclized library was 64,251, that is, the completed library, including nulls, contained 128,502 compounds. A partially single-stranded tag containing both PCR (polymerase chain reaction) priming regions and a unique molecular identifier (UMI) was added to enable PCR amplification and removal of duplicate sequences resulting from PCR, respectively.

DNA Encoded Library (DEL) Chromatographic Partitioning (Separation) and Sequencing of Fractions Produces Sequencing Count Chromatograms Liquid chromatography mass spectrometry (LCMS) and ultraviolet detection (UV) can be used to identify accurately the retention times (RT, tR) of DNA bound substrates. The library was collected in 96 different fractions and the amount of library in each fraction was determined by quantitative polymerase chain reaction. The library showed broad elution profiles consistent with the hydrophobic nature of the encoded peptides, contrasting sharply with the narrow elution profile of a blank library containing the same permuted DNA sequence but lacking any pendant warheads. The less aggressive gradient (10-40% acetonitrile gradient in aqueous TEAA buffer) was selected for sequencing based on its improved separation of late-eluting compounds and more uniform distribution of molecules across the fractions. Fractions were PCR-amplified using unique next-generation sequencing (NGS) adapters to encode each fraction. The encoded fractions were combined and sequenced in a single NextSeq run that produced 1.2 x 109 reads. UMI-based deduplication yielded a total of 5.1×10^8 unique reads across all fractions.

Because of the relatively even elution of the library throughout the run, equivolume aliquots from each fraction were taken into polymerase chain reaction (PCR) and were all run for the same number of amplification cycles. Each fraction was individually barcoded with unique iP7 adapters during the PCR in preparation for high throughput sequencing. The barcoded fractions were combined and sequenced all together in a Next Seq 2000 run with one billion reads. Because of the 1: 1 linear:cyclized mixture of the library, the overall theoretical library diversity was 128,502, which resulted in potentially 81 unique sequencing reads per compound per fraction (assuming a perfectly even distribution of each compound in each fraction).

The total sequencing reads were 7.6×10^8 counts (maximum of 1.2×10^9) and the deduplicated sequencing reads were 5.1×10^8 counts (67% unique reads). Additionally, the average total deduplicated (unique) sequencing counts per fraction were $5.3 \times 10^6 \pm 1.2 \times 10^6$, indicating good representation of each fraction in sequencing. The deduplicated sequencing counts of each compound in each fraction were determined. Generally, every compound in the library was observed in all fractions, with each molecule having an average median sequencing count in each fraction of 25.4, with fractions earlier in the run having slightly higher counts. However, each fraction contains significant sequencing count outliers. These large spikes in sequencing count are indicative of compound elution, with the average maximum sequencing count of each molecule in any fraction being 247.8, an approximately 10-fold ratio above baseline elution. Initial analysis revealed that 48% of the compounds showed identical elution profiles for their linear and cyclized versions, suggesting cyclization failure for these library members. Further investigation revealed

that compounds containing either an N-methyl or peptoid residue at the N-terminus failed to couple with azido acetic acid, preventing cyclization. After filtering these compounds, 30,720 cyclized sequences remained for detailed analysis.

Null-Informed Model to Guide Retention Time Assignment and Profile Reactivity To rigorously assess synthetic fidelity across the entire library, we developed an algorithm for comparing observed peaks to associated null-encoded truncation products. The inclusion of chemical truncations, encoded with a unique set of null DNA barcodes (N) at each synthesis cycle, ensured that all singly and doubly truncated synthetic failures were present in the library and were encoded with the same length of DNA as that of a complete molecule (CM). Comparison of the LC-seq trace of an expected CM with those of its corresponding nulls enabled identification of the expected product and its associated synthetic truncations. For example, for the library member cyclo[azidoAc-Leu(NMe)-N-isobutylGlyPro-Phe-Leu-Prg], consisting of the three building blocks BB55, BB03, and BB01 , a major late-eluting peak was observed in the CM chromatogram, which we assigned to the expected heptameric macrocycle. The CM chromatogram also contained minor peaks corresponding to the two single null truncations indicating loss of either BB03 or BB55. Each of these encoded null truncation chromatograms also contained peaks that aligned with associated double null truncations, providing additional confidence in the identification of these products.

The LC-seq chromatogram of the closely related CM cyclo[azidoAc-Val-Phe-Leu-Prg] contains a major peak which perfectly aligns with the single null truncation comprised of BB01 and BB03, indicating a failed coupling to BB05 in the third cycle of chemistry. After filtering out the uncyclized compounds, 10,007 compounds (33%) contained major peaks that matched truncation products, indicating synthesis failure, while 20,643 compounds showed unique, nonnull-matching major peaks, suggesting successful synthesis for these library members. The analysis revealed BB-specific trends in the chemistry. For example, peptoid BB11 showed poor coupling as a C terminal nucleophile with all BBs except the two β-homoamino acids (BB09 and BB10) likely due to reduced steric hindrance of these active esters. The less sterically hindered peptoid, BB12, showed higher coupling efficiency as the N-terminal nucleophile in the second synthetic step.

Library-wide Correlation with Drug-like Properties. The library was designed so that each scaffold contains multiple R-group variants, leading to a "liposcan" for each scaffold. There was a strong correlation between LC-seq retention times and calculated lipophilicity (AlogP) within a given scaffold, as well as significant differences in retention time among different scaffolds at the same ALogP, consistent with the results from the model compounds and providing further validation that the library was decoded correctly.

For example, two scaffolds that vary in stereochemistry at a single residue (D-Leu vs. L-Leu) showed significantly different retention times among matched pairs with the same combination of R groups;

however, within the same scaffold the retention times were highly correlated with the ALogP values of the individual liposcan variants. To further confirm the sequencing accuracy of the filtered LC-seq chromatograms and test whether LCseq retention times reflect membrane permeability in more biologically relevant assays, we individually resynthesized 24 compounds spanning different ring sizes and cleaved them from the solid support as their C-terminal ethyl esters.

The on-DNA retention times showed excellent correlation with off-DNA measurements across all ring sizes ($R^2 = 0.936$), confirming the LC-seq-assigned structural identities of these compounds (Fig. 4D). We then evaluated the passive membrane permeabilities of the 24 resynthesized compounds using the parallel artificial membrane permeability assay (PAMPA). While calculated AlogP values showed poor correlation with permeability, particularly for stereoisomers, on-DNA retention times showed strong, ring size-dependent correlations with PAMPA permeability. The permeability gain per unit increase in lipophilicity remained consistent across 5-mers, 7-mers, and 9-mers, with the 11-mers showing reduced gains, likely approaching size limitations for passive membrane permeability. As expected, the lipophilicity threshold required for high permeability increased with ring size, consistent with known MW-permeability relationships. The PAMPA measurements showed good agreement with MDCK cell permeabilities, supporting broader relevance for cellular penetration. These results demonstrate that the LC-seq platform provides not only quality control data but also biologically relevant insights into compound properties, enabling more informed DEL optimization.

Implications for extension to other drug-like properties. We introduce a transformative chromatographic strategy that enables direct assessment of both synthetic quality and drug-like properties for individual members of DNA-encoded libraries. This approach addresses two critical challenges in DEL-based drug discovery: identifying false positives from failed synthesis and predicting membrane permeability prior to resource-intensive resynthesis. The method's ability to track reaction efficiency for all building block combinations provides unprecedented insight for library design, while empirical measurements of lipophilicity enable rational prioritization of compounds for development.

Although demonstrated here with a focused 120,000-member library, the results suggest that the approach is scalable to million-member libraries with current sequencing capabilities and would be applicable to DNA-encoded libraries of small molecules in addition to peptidic macrocycles. We are investigating whether LC-seq can be applied to the analysis of mRNA display libraries, and more generally, to encoded libraries containing more polar and chemically complex side chains. The LC-seq platform could be adapted to evaluate additional drug-like properties through strategic modifications of the separation conditions. For metabolic stability assessment, the library could be exposed to liver microsomes or other specific metabolizing enzymes (e.g., proteases) before chromatographic analysis, with changes in retention time and peak profiles indicating metabolic modifications. Sequential time-point

analysis would enable quantitative stability measurements across the library. The large-scale property measurements enabled by LC-seq could dramatically enhance AI models' predictive capabilities, allowing more accurate evaluation of the vast virtual chemical space encompassed by larger macrocycles. This technology has immediate applications in library design and hit triage, potentially transforming the efficiency of DEL-based drug discovery.

APPLICATIONS

Drug discovery

New compound libraries including macrocycles and cyclic peptides

Quality control of DEL.

ADVANTAGES

Model that simultaneously assesses quality control of synthesis and predicts drug like properties of DNA encoded libraries of compounds.

INTELLECTUAL PROPERTY INFORMATION

Patent Pending

RELATED MATERIALS

▶ Massively parallel, single-molecule assessment of synthetic fidelity and drug-like properties in a DNA-encoded library - 07/14/2025