Quantitative codon optimisation of DNA libraries encoding sub-random peptides: design and characterisation of a novel library encoding transmembrane domain peptides

Ola Larsson, Dorit Thormeyer, Arian Asinger, Björn Wihlén, Claes Wahlestedt and Zicai Liang*

Center for Genomics and Bioinformatics (CGB), Karolinska Institutet, 171 77 Stockholm, Sweden

Received June 25, 2002; Revised September 16, 2002; Accepted October 9, 2002

ABSTRACT

Codons for amino acids sharing similar chemical properties seem to cluster on the genetic codon table. Such a geographical distribution of the codons was exploited to create chemically synthesised DNA that encodes peptide libraries containing only a subset of the 20 natural amino acids. The frequency of each amino acid in the subset was further optimised by quantitatively manipulating the ratio of the four phosphoamidites during chemical synthesis of the libraries. Peptides encoded by such libraries show a reduced complexity and could be enriched in peptides of a desired property, which are thus more suitable when screening for functional peptides. Proof of concept for the codon-biased design of peptide libraries was shown by design, synthesis, and characterisation of a transmembrane peptide library that contains >80% transmembrane peptides, representing a 160-fold enrichment compared with a fully randomised library.

INTRODUCTION

Currently, and in the past, researchers have considered all possible combinations of amino acids, initially by site-directed mutagenesis (1,2) and partial randomisation (3,4), as means for functional annotation of known genes as well as screening methods for peptides with novel properties. The 20 amino acids enable the creation of peptide libraries with very high complexity even when only short stretches of randomised regions are included. Although such libraries should contain peptides of a desired property, the specific peptide population may become extremely diluted, and screening for the peptides of interest may become difficult. One question is whether such a high degree of complexity is necessary in an artificial peptide library? Compared with the astronomical number of possible combinations of the 20 natural amino acids (20^{40} species for peptides with an average size of 40 kDa), the number of natural proteins (100 000 in human cells) is just a fraction. Although the bias towards this fraction of peptides has evolutionary reasons, it nevertheless argues for fully functional systems with limited complexity. Aptamer screening, using random nucleic acids or ribonucleic acids, has demonstrated that an affinity tag can be obtained for almost any bio-molecule from a library composed of only four different nucleotide building blocks (5). Can this be done using peptide libraries of similar complexity?

Peptide libraries can be generated either by direct chemical synthesis of polypeptides or by using peptide-encoding nucleic acid libraries. DNA-encoded peptide libraries are being employed more and more due to their compatibility with peptide panning technologies such as phage display, bacterial display and ribosome display. The philosophy of DNA (mostly oligonucleotide) library construction is still heavily dominated by the concept of total randomisation using all 20 amino acids. However, more limited randomisation involving fewer amino acids is widely used in site-directed optimisation (maturation) during peptide screening. In spite of numerous successes in the application of fully randomised libraries, two obvious drawbacks are the abundance of truncated peptides (due to stop codons) and the conflict between the theoretical library complexity and the physical complexity of the libraries (10^{11} for high-quality phage display libraries and 10^{12}–10^{13} for ribosome display libraries).

We have started to explore the concept of dissecting the totally randomised peptide library into a series of libraries that each contain only a subset of the amino acids at a predefined frequency. By doing so, we wish to transform the enormous unstructured fully randomised peptide library into a set of sub-libraries, well structured in terms of a desired property e.g. hydrophilicity, net charge, polarity or side chain size. In this paper we have designed a series of DNA-encoded peptide libraries containing a high percentage of transmembrane (TM) peptides as a proof of concept. Transmembrane receptors are typically anchored to the plasma membrane by a peptide stretch of 15–30 amino acids. These amino acids could function as an anchoring part of the receptor. However, recent data suggest that the membrane part of the receptor is important for intra- and inter-molecular interactions (6–13),

*To whom correspondence should be addressed. Tel: +46 8 7286371; Fax: +46 8 327934; Email: zicai.liang@cgb.ki.se
e.g. it is known that interactions between the seven TM domains in G-protein coupled receptors (GPCRs) are important for receptor dimerisation and signalling. It has also been shown that peptides that perturb the interactions between the TM domains disrupt the dimerisation process and interfere with GPCR function (14–16). Thus, it is likely that the membrane part of the receptor has more functions than just anchoring the receptor to the membrane, e.g. mediation of specific interactions or conformational changes. Our TM peptide libraries can be used to screen for peptides that can be used for intervention of membrane–protein interactions.

MATERIALS AND METHODS

The natural clustering of the codons

Codons for the same amino acid (apart from serine and arginine) differ from each other only in the third nucleotide of the triplet (Table 1). Codons are, however, not only clusters for single amino acids. In our assessment of the codon table (Table 1), we observed that amino acids with similar properties seem to cluster to the same location of the codon table. One striking example is that the first column of the table contains only hydrophobic amino acids. Based on these phenomena, we have designed a series of so-called codon-biased nucleotide libraries that encode random peptides, which contain only a subset or a biased selection of the 20 natural amino acids

Library construction

TM library 1, TM library 2, TM library 3 (Table 2) (BioAsia, Shanghai, China) and a randomised library [NNN]$_{18}$ (Interactiva, Germany) contain the library part flanked by a 5\textquotesingle start codon followed by an N-glycosylation signal (ATG AAT GCT AGT) and a 3\textquotesingle sequence harbouring a BamHI restriction site (GGA TCC TAT GCA CC). The yield from the synthesis of the TM libraries after PAGE purification was ~5 nmol. This indicates a library size of $3 \times 10^{14}$. The construct was PCR amplified using primers 5\textquotesingle-C ACC TAA AGA ATG AAT GCT AGT and 5\textquotesingle-C AAC TGA ATT AGG TGC ATA GGA to select for full-length oligos [94°C 1 min, 39°C 1 min, 72°C 30 s]$_{2}$ (94°C 30 s, 55°C 30 s, 72°C 30 s)$_{12}$ 72°C 4 min]. The PCR selected library was cloned into the pcDNA 3.1D/V5-His-TOPO vector (Invitrogen, Sweden).

Library sequencing

We sequenced TM library 3 clones to verify that the encoding of peptides can be accurately synthesized as designed. We also sequenced clones from the randomised library to generate negative controls. DH10b cells (Life Technologies, Sweden) were transfected with the library and individual colonies were PCR amplified using the T7 primer (5\textquotesingle-TAA TAC GAC TCA CTA TAG GG) and the BGH reverse primer (5\textquotesingle-GAG CTA CAG TCG AGG). Full-length inserts were selected on an agarose gel for sequencing, which was performed using DYEnamic ET Terminator Cycle Sequencing (Amersham Pharmacia, Sweden) and the T7 primer. Sequences were analysed using DNAtools (www.dnatools.dk).

Cellular localisation

Thirteen TM peptides and three peptides derived from a fully randomised library were selected for localisation studies. These were PCR amplified with primers 5\textquotesingle-CGA TCA GAG AAT TCA TGA ATG CTA GT-3\textquotesingle' and 5\textquotesingle-GCT GGT GCA TAG GA-3\textquotesingle' and then cloned into pEGFP-N3 (Clontech, USA) using BamHI and EcoRI sites. The TM peptides expressed from these constructs were all attached to the N-terminus of the recombinant green fluorescent protein (GFP). Constructs were sequence verified. The three random peptides were the only stop codon-free peptides out of 20 clones sequenced from a fully randomised library. We did not include additional random peptides as negative controls as it was demonstrated that only ~0.1% of such random peptides could have a membrane localisation (17). We did not include the other 17 peptides from the random library in this assay because they would simply have abolished the GFP translation. GFP was used as a negative control and the $\alpha_{2c}$-adrenergic receptor, fused to GFP, was used as a positive control. All constructs were transfected into Cos7 cells or 293
cells using Trans Fast transfection reagent (Promega, USA) and harvested after 24 h. The Cos7 cells were permeabilised (0.5% Triton X-100, 20 mM Tris–HCl pH 7.4, 50 mM NaCl, 300 mM sucrose, 3 mM MgCl₂) for 1 min, washed twice in PBS, fixed in 3.7% formaldehyde (PBS) for 5 min, washed three times in PBS and mounted onto glass slides. The 293 cells were washed once in PBS and mounted. The cellular localisation of the GFP-coupled peptides was observed using a fluorescence microscope (Leica DMRXA).

RESULTS AND DISCUSSION

Library design using the codon bias approach

By exploiting the fact that amino acids that share some bases in their triplets, and have similar properties, form clusters in the codon table (Table 1), it is possible to design encoded peptide libraries with different properties using codon-biased library design. This is done by defining each base of the triplet as a mixture of all bases: (TA%CB%AC%GD%) where A + B + C + D = 100%. The different proportions of the bases are mixed and used during the conventional synthesis of the oligo so that a biased randomness is created. Figure 1 shows a subset of libraries that can be constructed: e.g. a library containing hydrophilic negatively charged peptides (50% D and 50% E) could be constructed by synthesising a library with the formula [GAN]ₙ, meaning that the first base in the triplet is (G₁₀₀%), the second is (A₁₀₀%) and the third is (A₂₅%T₂₅%G₂₅%C₂₅%). In a similar way many different libraries with varying properties can be designed.

Construction and evaluation of a TM domain peptide library

Table 2. By using codon-biased library design, which is achieved during standard oligonucleotide synthesis, random libraries with a bias towards some amino acids can be generated.

<table>
<thead>
<tr>
<th>Formula of the library</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[(T₂₇%C₂₀%A₂₆.₅%G₂₆.₅%)(T₆₅%C₂₀%A₅%G₁₀%)(T₃₀%C₅%G₄₀%)]₁₈</td>
<td>TM library 1</td>
</tr>
<tr>
<td>[(T₃₀%C₂₅%A₂₇.₅%G₂₅%)(T₇₀%C₁₅%A₅%G₁₀%)(T₃₀%C₅%G₄₀%)]₁₈</td>
<td>TM library 2</td>
</tr>
<tr>
<td>[(T₂₇%C₂₀%A₂₆.₅%G₂₆.₅%)(T₆₈%C₂₀%A₅%G₁₀%)(T₃₀%C₅%G₄₀%)]₁₈</td>
<td>TM library 3</td>
</tr>
</tbody>
</table>

The table shows the design of three TM libraries. Each was generated to improve the stop codon frequency, the percentage of 18mers classified as membrane compatible and the amino acid distribution of the previous one.

Figure 1. By using codon-biased library design many different libraries encoding a subset of the 20 amino acids can be generated. The figure shows miniaturised codon tables (same design as Table 1) where one rectangle represents one codon. The amino acids that are selected in each library have been marked with black. Under each codon table the formula for generating such a library is shown. To generate a bias within the selected amino acids in the library the individual relative amounts of the four bases can be further modulated.
library design. The hydrophobic amino acids, which are highly abundant in natural TM domains, appear to form a group that has T as the second base in their triplet. Similarly, the charged amino acids, which are of low abundance and whose existence close to each other in a peptide drastically reduces the likelihood of the peptide to appear in the membrane (18), have A as their second base. Based on such observations three different libraries were designed with each one being a refinement of the previous one (Table 2).

The quality of the TM libraries was evaluated using three different criteria: (i) percentage of the encoded peptides (18mers) with stop codons; (ii) the occurrence of amino acid compositions that are defined as less likely to appear in a natural TM domain [Gromiha (18) concluded that an 18mer peptide was unlikely to be a TM peptide if it contained three amino acids out of lysine, glutamate, arginine, aspartic acid, glutamine, asparagine, histidine and proline]; (iii) how well the library mimics the relative amino acid abundance in naturally occurring TM domains.

Of the encoded 18mer peptides in a totally randomised library, 84% had stop codons in any position. This is improved 20-fold in TM library 3 (the proportion that contains stop codons is calculated as 4%). The occurrence of non-TM peptides, which are defined as peptides that contain three or more non-TM amino acids, was calculated to decrease from 97% in a randomised library to 10% in TM library 3. If all peptides that are either truncated or non-TM are regarded as non-functional, TM library 3 theoretically contains 86% functional peptides that can enter the membrane, in contrast to 0.5% for a totally randomised library (Table 3).

To verify that the chemically synthesised DNA library can be constructed according to the theoretical design, 50 independent clones from TM library 3 were sequenced. One of the clones was found to contain only 53 nt, instead of containing 54 nt according to the formula. This probably represents a truncated product from the chemical synthesis. Of the remaining 49 clones, three contained a stop codon and six encoded peptides that are unlikely to be TM peptides due to their amino acid composition (Fig. 3). Thus the experimental data suggested that ~82% of the clones from library 3 encode TM peptides, which is close to the expected frequency of 86%. Figure 2B shows the amino acid abundance of the 49 sequenced full-length encoded peptides, theoretically calculated values for TM library 3 and natural TM domains. The sequenced library displays similar amino acid abundance, as we predicted, and mimics the amino acid distribution in natural TM domains.

To evaluate the TM libraries we monitored the occurrence of stop codons and percentage of 18mers with a non-TM amino acid composition. The percentage of functional TM peptides within each library is defined by subtracting the peptides containing stop codons and non-TM composition from the total. The table shows the calculated relative amounts of encoded 18mer peptides with stop codons and encoded peptides classified as functional in a fully randomised library, the three codon-biased TM libraries and 49 sequenced encoded peptides from TM library 3.

### Table 3. Comparison of different libraries for TM peptide ratios

<table>
<thead>
<tr>
<th></th>
<th>Membrane compatible peptides (18mers) (%)</th>
<th>18mers with stop codon (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random library</td>
<td>0.5</td>
<td>84</td>
</tr>
<tr>
<td>TM library 1</td>
<td>71</td>
<td>10</td>
</tr>
<tr>
<td>TM library 2</td>
<td>73</td>
<td>11</td>
</tr>
<tr>
<td>TM library 3</td>
<td>86</td>
<td>4</td>
</tr>
<tr>
<td>Sequenced encoded peptides from TM library 3</td>
<td>82</td>
<td>6</td>
</tr>
</tbody>
</table>

In a complex example such as the TM library, where all amino acids need to be present but the individual relative amino acid occurrence has to be changed compared with a random library, the actual amino acid distribution is another way to evaluate the library. In (A) the theoretical relative amino acid abundance in a fully randomised library, the TM libraries and the natural TM domains is shown. The three libraries do not differ substantially from each other and they mimic the natural distribution of amino acids in TM domains well. (B) shows the relative amino acid abundance in 49 sequenced peptide-encoding library clones from TM library 3, the predicted relative amino acid abundance in TM library 3 and the relative amino acid abundance in natural TM domains.
Figure 3. To monitor that the sequenced peptides encoded by TM library 3 are membrane compatible, Kyte and Doolittle hydropathy plots of the 49 sequenced peptides from TM library 3 were made. Positive values in the plot indicate that the three amino acids in the sliding window are hydrophobic.
increased the occurrence of functional TM peptides in TM library 3 (10/13) compared with a fully randomised library (0/20).

CONCLUSION

Peptide ligands are important resources for functional proteomics and drug development. Different partially randomised schemes have been employed for constructing peptide libraries through chemical synthesis (19). By using codon-biased library design it is possible to formulate a systematic strategy to generate different peptide libraries with small subsets of all amino acids or a strong bias towards some amino acids with shared properties. Such libraries, encoded by chemically synthesised oligonucleotides and containing only a fraction of all possible peptides, may provide a shortcut for the screening of peptide ligands. Compared with the chemical synthesis of peptide libraries, the generation and re-generation of codon-biased libraries is much easier and very flexible. The construction of the library involves only standard molecular biology techniques, and the library can be integrated easily into phage display, bacterial display or ribosome display protocols for high throughput screening.

The main limitation of the codon-directed approach is that the flexibility, in terms of amino acid combination, is limited by the geographical distribution of codons in the codon table. This is reflected by either the unwanted incorporation of stop codons and other amino acids, or a partial loss of some amino acids (especially Trp due to the geographical linkage with the stop codons). Even with severe bias against Trp in the TM libraries, the abundance of functional TM peptides containing this residue is still higher in our libraries compared with a fully randomised library. This is mainly due to the fact that our library has an ~160-fold elevated abundance of TM peptides. Similar mathematics could apply to many of the codon-biased libraries.
ACKNOWLEDGEMENT

This work was supported by a grant from the Center for Genomics and Bioinformatics, Karolinska Institutet and Pharmacia Corporation to Z.L.

REFERENCES