Biochemistry

© Copyright 2002 by the American Chemical Society Volume 41, Number 17 April 30, 2002

Articles

A Combinatorial Approach toward Analyzing Functional Elements of the *Escherichia coli* Hemolysin Signal Sequence[†]

David Hui and Victor Ling*

British Columbia Cancer Research Centre, Vancouver, British Columbia V5Z 1L3, Canada, and Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

Received July 9, 2001

ABSTRACT: Secretion of hemolysin is directed by a signal sequence located within its C-terminal 60 amino acids. Deletion analyses have indicated that the extreme end of this C-terminus is critical for transport; however, it is not known if this region contains structural features necessary for function. In this study, we have used a combinatorial approach to generate two contiguous 8-residue random libraries (Cterm1 and Cterm2) in the signal sequence to investigate the functional specificity of the last 16 residues. The large number of variants generated had provided us with a rich data set to determine if a restricted subset of sequences was actually required for function in the extreme C-terminus. We observed that over 90% of the random sequences in the Cterm1 region were secreted at close to wild-type level, while the Cterm2 region was more restricted with only 50% of the random sequences supporting wild-type-like transport. It appeared that, in the Cterm2 region, the relative lack of positive charge is favored for function. These findings, along with previous results, allow us to propose a model for recognition and transport of hemolysin that emphasizes secondary structure and general biophysical properties over primary sequence. This model may have implications for understanding the broad substrate specificity common among ATP-binding cassette transporters.

In this study, we have investigated the substrate specificity of a well-characterized prokayotic ATP-binding cassette (ABC)[1] transporter, the *Escherichia coli* hemolysin system. Secretion of the 107 kDa hemolysin is directed by a C-terminal signal sequence located within the last 60 amino acids of this protein (*1*). Although the hemolysin transporter complex has been designed through evolution for the translocation of a single substrate (i.e., hemolysin), it demonstrates the remarkable ability to transport other proteins with drastically different primary sequences (*2*−*4*). Since many ABC proteins share a similar mechanism of transport (*5*), it is hoped that some of the principles derived from this study can be applied to further enhance our understanding of the broad substrate specificity of other members of the ABC transporter superfamily. The hemolysin system presents an excellent model for our substrate specificity investigation for a number of reasons. First, its prokaryotic nature allows rapid and reproducible experimentation. Second, the substrate is a protein encoded on a plasmid, which permits easy manipulation. Third, reliable assays have been established to measure secretion level, facilitating functional analysis.

[†] This work was funded by the Medical Research Council of Canada (Grant 14137). D.H. was supported by a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada.

* To whom correspondence should be addressed. Telephone: (604) 877-6152. Fax: (604) 877-6150. E-mail: vling@bccancer.bc.ca.

[1] Abbreviations: ABC, ATP-binding cassette; PCR, polymerase chain reaction.

10.1021/bi011425g CCC: $22.00 © 2002 American Chemical Society
Published on Web 04/02/2002

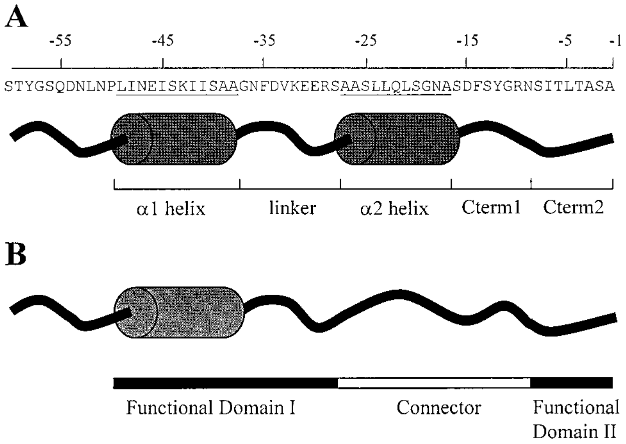

FIGURE 1: Structural and functional models of the *E. coli* hemolysin signal sequence. (A) The structural model of the hemolysin signal sequence is shown with its helix (α1)−linker−helix (α2) motif, with residues representing the two conserved helices underlined. In the current study, we used a combinatorial approach to generate two contiguous 8 amino acid random libraries (Cterm1 and Cterm2) to investigate the functional specificity of the last 16 amino acids. (B) A new functional model was proposed on the basis of analyses of all six combinatorial libraries [from this study and a previous one (*6*)]. Two functional domains (■) and one connector region (□) were identified. An α1 amphiphilic helix was found to be important in the first domain, while hydrophobicity is crucial in the second domain. A minimum of 4−7 residues is required in the connector region for efficient transport.

The specificity of hemolysin transport still has not been fully understood, however. Previous attempts to decipher the code within the signal sequence include numerous deletions and point mutations (*2*, *3*). These approaches achieved limited success because it is difficult to distinguish whether a reduction in secretion activity is due to the omission of a critical element or to the presence of the newly introduced abnormality. Thus, the signal sequence remains to be assessed in a more objective and systemic manner.

Recently, we have utilized a combinatorial approach to investigate the functional features within the conserved helix−linker−helix motif (−49 to −17, Figure 1A) of the hemolysin signal sequence (*6*). This technique can be divided into four sequential steps: (1) generation of a large data set of sequence variants in a specific region of interest, (2) determination of the activity of each variant with functional assays, (3) transformation of amino acid sequence information into biophysical data using predictive algorithms, and (4) correlation between the genotype (biophysical data) and phenotype (activity) to identify properties that enhance/prevent function. This approach has the advantage in that it allows drastic alterations to be engineered into a well-defined region while maintaining the natural spatial relationships of the target and nontarget regions. Furthermore, a large number of variants can be generated, allowing for a systematic analysis of functional elements with statistical confidence. Three contiguous random libraries (α1, linker, and α2) were generated in which the target regions were replaced by random sequences (Figure 1A). The ability to support transport was then measured in the random variants. Using this approach, we have made the surprising observation that the conserved α2 helix is not required for transport, while the amphiphilic nature of the α1 helix is a critical determinant of function. This work has demonstrated that a combinatorial approach, generating random sequence variants, can be highly informative of the functional necessities of even conserved protein structures.

In the present study, we have extended this combinatorial approach to investigate the extreme C-terminus of the signal sequence. Experience from previous studies has shown that this region is particularly challenging to analyze because of its seemingly unlimited capacity to tolerate almost any mutations, with the exception of large deletions (*7*, *8*). Two contiguous 8 amino acid random libraries (Cterm1 and Cterm2, Figure 1A) were created to explore the range and nature of primary sequences that can be accommodated by the hemolysin transporter system. The large number of signal sequence variants generated has provided us with a rich data set to determine if a restricted subset of sequences is actually required for function in the extreme C-terminus. We have made the interesting finding that the Cterm1 region has no structural requirement, while the Cterm2 region favors nonpositively charged residues for transport. This study has also demonstrated the application of combinatorial analysis as a powerful tool for defining the limits of an extremely versatile biological system.

EXPERIMENTAL PROCEDURES

Bacterial Strains and Plasmids. The *E. coli* strains Top10F′ (F′{*lac*I^q, Tn*10*(Tet^R)} mcrA Δ(*mrr-hsd*RMS-*mcr*BC) *ϕ*80 *lac*ZΔ*M*15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L(Str^R) *end*A1 *nup*G) (Invitrogen) and JM83 (*λ*$^-$ *ara*Δ(*pro-lac*) *rps*L *thi* *ϕ*80 *dlac*ZΔ*M*15*λ*$^-$) were used for all cloning procedures. pUCAC494 and pUCAC494BN both encode hemolysin (*6*) and were used for the generation of the Cterm1 and Cterm2 random libraries, respectively. All pUCAC494 variants were selected with ampicillin at 50 *μ*g/mL. *E. coli* JM83 harboring pLGBCD (*6*) was transformed with mutated versions of pUCAC494 and assayed on blood agar plates. pLGBCD, which encodes the transporter genes (*hlyB*, *hlyD*) and toxin activation gene *(hlyC*), was selected with chloramphenicol at 34 *μ*g/mL.

Cloning of Random Library Mutants. The procedure for random oligonucleotide mutagenesis was as described previously (*6*). A list of the cloning plasmids, restriction sites, and oligonucleotide sequences for each random library is shown in Table 1. All random variants were selected with colony polymerase chain reaction (PCR). The screening of Cterm1 mutants was based on the idea that random oligonucleotide mutagenesis would likely destroy many of the previously mapped restriction sites within the targeted region. Therefore, amplification of a PCR fragment and subsequent digestion with a selected restriction enzyme would allow the selection of clones with a PCR product that could not be cleaved and thus likely contained a random sequence. A 579 bp fragment was amplified from each colony using primers HA24 (5′-GATTTCCGGGACGTTGCC-3′) and M13R1 (5′-AAAACGACGGCCAGTGAATTC-3′) and then subjected to digestion with *Nde*I. The Cterm2 region did not contain any convenient restriction sites, and thus a second method of screening was adopted. Primers HA24 and Cterm-test-R (5′-GCTGATGCTGTCAAAGTTATTG-3′) were used to amplify a 470 bp fragment. Cterm-test-R was designed to anneal to the wild-type Cterm2 region only. Therefore, absence of a PCR band would indicate a positive candidate.

Table 1: Sequence of Oligonucleotides Used To Assemble Combinatorial Cassettes[a]

Cterm1 cassette: inserted into pUCAC494 at the BglII and KpnI sites	
HlyA-Cterm1-F	5'-GAAAGATCTGCCGCTAGCTTATTGCAGTTGTCCGGTAATGCCNNNNNNN NNNNNNNNNNNNNNNNNNNNTCAATAACTTTGACAGCATCAGCATAATATATTA-3'
HlyA-Cterm1-R	5'-GAGCTCGGTACCATTATGACTCCAAAAAAATAGCAATCTTATGTGGCAC AGCCCAGTAAGATTGCTATCATTTAAATTAATATATTATGCTGATGCTGTC-3'
Cterm2 cassette: inserted into pUCAC494BN at the NheI and KpnI sites	
HlyA-Cterm2-F	5'-CTGCCGCTAGCTTATTGCAGTTGTCCGGTAATGCCAGTGATTTTTCATAT GGACGGAACNNNNNNNNNNNNNNNNNNNNNTAATATATTAATTTAAA-3'
HlyA-Cterm2-R	5'-GAGCTCGGTACCATTATGACTCCAAAAAAATAGCAATCTTATGTGGCAC AGCCCAGTAAGATTGCTATCATTTAAATTAATATATTA-3'

[a] Underlined region represents complementary sequence where annealing occurred.

Although this procedure did not require digestion, a follow-up PCR reaction using HA24 and M13R1 was used to eliminate any candidates with abnormal band size or no band at all.

All positive colonies were grown overnight, and the plasmid DNA of each variant was obtained with the Quantum miniprep kit (Bio-Rad). Samples were prepared for sequencing with the ABI PRISM BigDye terminator cycle sequencing ready reaction kit according to manufacturer's instructions and analyzed on a 310 Genetic Analyzer (PE Biosystems). The forward sequence was obtained for all variants using αSEQ (5′-GACGGCAGGGTAATCACACC-3′) as primer. The reverse sequence of selected variants was obtained using M13R1.

Blood Agar Plate Assay. The hemolysin secretion level of each variant was determined using a blood agar plate assay similar to that of a previous study (*6*). Briefly, plasmid DNA encoding hemolysin variants was transformed into JM83 containing pLGBCD and spread on blood agar plates (bottom agar, 10 mL of LB; top agar, 10 mL of LB with 5% sheep blood) in triplicate. After an incubation period of 19 h at 37 °C, each variant was assigned a zone rank from 0 (no hemolysis) to 6 (wild type) by comparison to a set of standards. All plates were examined twice separately in a blinded fashion. Altogether, six readings were taken for each variant. The average and standard deviation were calculated to provide an indication of the secretion level and variability, respectively. The percentage of secretion relative to wild type for each rank as determined by enzyme-linked immunosorbent assay was approximately rank 6 = 100%, rank 5 = 90%, rank 4 = 50%, rank 3 = 30%, rank 2 = 10%, rank 1 = 2%, and rank 0 = 0%. In this study, a hemolytic zone assignment of 4 or above was considered to be efficient transport.

Since an average of 100 variants were assayed for each library at the same time, the blood agar plate assay proved to be the only feasible method for phenotype determination because it is quick and convenient. However, this procedure is not completely ideal in that it is semiquantitative and the hemolytic zone assignments are determined by eye and therefore subject to personal bias. To determine the reproducibility of this assay, a second person was asked to rank all Cterm1 random variants (by comparison to 7 standards) independent of the primary observer. All readings were compiled, and the two sets of data were analyzed for differences (data not shown). Out of a total of 98 variants, 21 were given the same assignments as before. The Pearson product moment correlation coefficient between the two sets of readings was 0.95. This level of correlation was considered acceptable for the purpose of this study. Furthermore, selected variants from each of the two random libraries were plated together in a second round to facilitate a direct comparison. The Pearson correlation coefficient between the first and second sets of measurements for the two libraries was 0.93 (data not shown), confirming the high reproducibility of the blood agar plate assay.

Data Analysis. DNA sequences were translated and analyzed with the Wisconsin Package Version 9.1, Genetics Computer Group, Madison, WI. Additional Perl scripts were written by Dr. Eric Cabot and D. Hui to facilitate bulk analysis.

SDS−Polyacrylamide Gel Electrophoresis and Western Blotting. To determine the amount of endogenous hemolysin, JM83 bacteria were transformed with plasmids encoding selected hemolysin variants (stop, random, or wild type) and harvested at $OD_{600} = 0.85 \pm 0.05$. Following centrifugation (7000*g* for 15 min), cell pellets were resuspended in STE buffer (10 mM Tris, pH 8.0, 150 mM NaCl, 1 mM EDTA) supplemented with various protease inhibitors. For each sample, an equivalent of 200 μL of cells was boiled for 3 min and run on an SDS−polyacrylamide gel (7.5% separating) under reducing condition. After the standard Western transfer and blotting procedure [with anti-hemolysin antiserum at 20000× dilution and goat anti-rabbit antibody at 10000× dilution (Jackson ImmunoResearch Laboratories, Inc.)], visualization was achieved by using the enhanced chemiluminescence Western blotting detection reagents (Amersham Pharmacia Biotech). It was found that the majority of the combinatorial mutations in the signal sequence did not have a dramatic effect on the quantity of intracellular hemolysin as compared to wild type (data not shown). Thus, the hemolytic zone size as seen on blood agar plate assays would be a good indicator of secretion level.

Systematic Deletion of the Connector Region. Internal deletions of 3, 6, 9, 12, 15, and 19 residues were engineered in the connector region (linker and α2), all of which started at position −27. PCR fragments were amplified from pUCAC494 using primers M13R1 and one of HlyA-del3-F (5′-GGAAGATCTTTATTGCAGTTGTCCGGTAAT-3′), HlyA-del6-F (5′-GGAAGATCTTTGTCCGGTAATGCCAGTG-3′), HlyA-del9-F (5′-GGAAGATCTAATGCCAGTGATTTTTCATAT-3′), HlyA-del12-F (5′-GGAAGATCTGATTTTTCATATGGACGGAAC-3′), HlyA-del15-F (5′-GGAAGATCTTATGGACGGAACTCAATAACT-3′), and

Table 2: Observed Nucleotide Frequency within Different Random Libraries

nucleotide	Cterm1 random library (%)	Cterm2 random library (%)
T	30	12
C	19	26
A	28	14
G	23	48
sample size[a]	1992	1800
p value[b]	1.4×10^{-12}	1.5×10^{-130}

[a] The sample size equals the total number of random and stop variants multiplied by the number of nucleotides in each target region (i.e., 24). [b] χ^2 value was calculated with expected ratio = 25% for each nucleotide, and the *p* value was determined with degrees of freedom = 3.

HlyA-del19-F (5′-GGAAGATCTTCAATAACTTTGACAGCATCA-3′). The mutated fragments were introduced back into wild-type plasmid at the *Bgl*II and *Kpn*I sites. Positive clones were subjected to sequencing as outlined above, and the phenotype of these mutants was determined by blood agar plate assay.

RESULTS

Nucleotide and Amino Acid Distributions of Random Library Variants. The extreme C-terminus of the hemolysin signal sequence, defined here as the last 16 amino acids, was analyzed by random oligonucleotide mutagenesis. For improved resolution, two contiguous 8 amino acid random libraries were generated. The boundaries for these libraries were decided upon on the basis of a previous study suggesting that the last 8 amino acids constitute a functional region (*8*). The resulting mutants had a string of random nucleotides (G, A, T, C) in place of the wild-type target region. Upon isolation of the variants, the genotype was determined by sequencing of the entire secretion signal. Each mutant was classified into one of three classes on the basis of the DNA sequence: those of the intended design (simply called random mutants), those that had a premature stop codon (stop mutants), and those that contained unanticipated mutations elsewhere in the signal sequence (not analyzed). For each of the Cterm1 and Cterm2 libraries, 20^8 different random variants and $\sum 20^x$ (where $x = 0-7$) distinct stop variants could be generated theoretically.

To obtain an idea of the randomness of the resulting libraries, the nucleotide frequency within the mutated region was determined for all random and stop variants of each library. Although the random regions were designed such that all four nucleotides had an equal chance of being incorporated at each position (i.e., 25%), the actual proportion deviated dramatically (Table 2). On the basis of χ^2 statistics, it was determined that these frequencies could not have arisen by chance. This skewing could be a result of manipulations during library generation and/or biological selection.

Previous analysis of other hemolysin signal sequence random libraries (*6*) suggested that biological selection is a less likely explanation for this skewed distribution. This was because the observed amino acid frequency closely matched the frequency calculated from the observed nucleotide ratio within each library. When χ^2 statistical calculation was performed on the Cterm1 and Cterm2 random libraries, *p* values of 8.7×10^{-5} and 4.4×10^{-1} were obtained, respectively (Table 3). The calculated result for the Cterm2 library is consistent with previous findings; however, the relatively small *p* value of the Cterm1 random library indicated that there is a very small chance of obtaining the observed amino acid frequency based on the observed nucleotide frequency. While we do not understand the reason for this deviation, all amino acids are represented in significant proportions in both libraries, providing a diverse population of combinatorial mutants for further analysis.

Table 3: Amino Acid Frequency within Different Random Libraries

	Cterm1 random library		Cterm2 random library	
amino acid	obsd (%)[a]	predicted (%)[b]	obsd (%)[a]	predicted (%)[b]
G	5.17	4.22	23.47	23.00
A	4.37	4.37	12.38	14.50
V	6.85	6.02	6.00	5.00
L	10.39	8.43	4.12	5.67
I	6.50	5.27	0.87	0.83
P	3.70	2.86	6.53	5.67
F	4.48	6.33	0.58	0.17
W	1.56	0.90	2.91	4.17
Y	4.15	2.86	0.64	1.00
T	5.37	6.48	3.48	3.83
S	8.92	11.60	5.67	5.33
C	3.38	5.72	2.28	2.17
M	1.91	1.81	0.82	1.00
Q	2.72	2.26	2.16	1.83
N	3.84	4.82	0.70	0.17
D	3.13	2.56	2.50	2.33
E	3.22	1.66	4.09	2.83
K	3.95	6.02	1.15	1.50
R	7.59	7.83	16.47	16.00
H	2.65	1.51	1.32	1.17
stop codon	6.17	6.48	1.86	1.83
sample size[c]	664		600	
p value[d]	8.7×10^{-5}		4.4×10^{-1}	

[a] The observed amino acid frequencies were determined for all random and stop variants within each library. [b] The predicted amino acid frequencies were calculated on the basis of the observed nucleotide frequencies from Table 1. [c] The sample size equals the total number of random and stop variants multiplied by the number of amino acids in each target region. [d] The χ^2 value was calculated on the basis of the observed and predicted frequencies from above, and the *p* value was determined with degrees of freedom = 20.

Secretion Efficiency of Cterm1 and Cterm2 Random Variants. Forty-eight random variants were generated in the Cterm1 region to determine its functional role. If this region contains critical structural element(s), most Cterm1 random mutants will be expected to be secreted at low levels since they will not contain the specific structural feature. The hemolytic zones for Cterm1 random mutants ranged from 2.8 to 6. The population was heavily skewed to the right, with a mean of 5.6 and a median of 6 (Figure 2A). Forty-four of the 48 random mutants secreted at 5 or higher (i.e., close to wild-type level). This distribution showed that the Cterm1 region can tolerate almost any combination of amino acids without having a major effect on transport, providing strong evidence that this region contains no primary or secondary structural elements required for efficient transport.

The second random library, Cterm2, involved replacing the last 8 amino acids of the signal sequence with a degenerate sequence. The hemolytic zones for the 65 Cterm2 random variants ranged from 1 to 6. Interestingly, a bimodal distribution was observed (Figure 2B). About half of the population (32 out of 65) was secreted at efficient levels (with a hemolytic zone rank ≥4), while the rest was secreted at suboptimal levels. The extreme C-terminus must contain

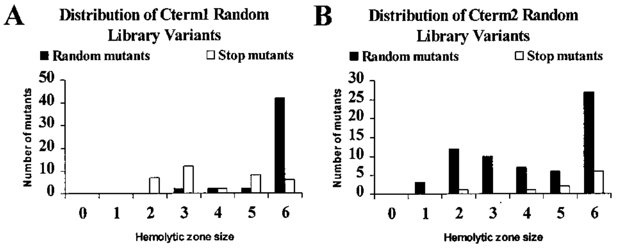

FIGURE 2: Distribution of Cterm1 and Cterm2 random library variants. Upon sequencing, the phenotype of each mutant was determined by blood agar plate assay. On the basis of the size of the halo, each variant was assigned a rank from 0 (no hemolytic zone) to 6 (wild type). The secretion patterns of variants from (A) the Cterm1 random library (48 random and 35 stop) and (B) the Cterm2 random library (65 random and 10 stop) are plotted above. Each secretion category has a 0.5 margin (e.g., a hemolytic zone size of 5 represents mutants secreting between and including 4.5 and 5.5).

Table 4: Kyte−Doolittle Hydrophilicity of Combinatorial Variants[a]

	range	average[b]	correlation[c]
Cterm1 random	−1.52 to 1.06	−0.19	−0.24
Cterm2 random	−0.81 to 1.93	0.28	−0.71[d]

[a] The Kyte−Doolittle hydrophilicity was determined with Wisconsin Package Version 9.1, Genetics Computer Group (GCG). A window of 7 residues was used, resulting in a value for each amino acid within the mutated signal sequence. [b] The hydrophilicity values for residues within the mutated region of each mutant were averaged. Following that, a mean was obtained for each library. [c] The Spearman correlation coefficient between the average hydrophilicity value and hemolytic zone size was determined for each library. [d] Significant at the 0.01 level (two tailed).

some element critical for transport since secretion was severely affected in half the population. Moreover, the observation that about 50% of the Cterm2 random variants were secreted at close to wild-type levels suggests that this critical element can, nevertheless, be satisfied by many random combinations of amino acids and is thus likely to be something general.

In an attempt to actually identify the functional feature(s) involved in the Cterm2 region, the primary sequences of the random variants were manually inspected. It was found that variants with more than one positive charge in the last 8 amino acids were generally secreted at lower levels. Thus, efficient transport requires the relative lack of positively charged residues in the extreme C-terminus. When the Kyte−Doolittle hydrophilicity was calculated for the mutated region of each Cterm2 random mutant and analyzed in the context of transport efficiency, a Spearman correlation coefficient of −0.71 ($p < 0.01$) was obtained (Table 4). This correlation suggested that hydrophobicity in the Cterm2 region likely plays an important role in transport efficiency. However, it was recognized that the Cterm2 library has an abundance of lysine residues, and thus the hydrophobicity element could simply be a reflection that the signal sequence cannot tolerate too many positive charges. No trend was observed when the same analysis was performed for Cterm1 variants (Table 4). This was not surprising since efficient secretion was obtained regardless of the primary sequence in the Cterm1 region.

Secretion Efficiency of Cterm1 and Cterm2 Stop Variants. The stop mutants obtained from both random libraries were also analyzed. These variants were those that contained a stop codon in the targeted region, resulting in truncation of the downstream sequence. Thirty-five Cterm1 stop variants were generated. In these mutants, the last 16 amino acids were replaced by a random combination of 0−7 amino acids. Their secretion levels ranged from 2 to 6. A bimodal distribution similar to that of Cterm2 random variants was observed. This pattern was clearly illustrated when the hemolytic zone was plotted against the position of the stop codon (Figure 3) as most variants fall into either the top portion or the bottom portion of the graph. Interestingly, a significant proportion of Cterm1 stop mutants was still secreted at wild-type level despite a truncation of 8 or more residues! When the length element was controlled by comparing mutants with a stop codon at the same position, those that contained positively charged residues between positions −2 to −8 were consistently transported at levels lower than their counterparts, confirming the finding from the Cterm2 random library. The zone size for the 10 Cterm2 stop mutants ranged from 2 to 6, with a mean of 5.2 and a median of 5.7. Since the number of mutants in this category was small, no extensive analysis was performed. However, a bimodal distribution would be expected had more mutants been generated.

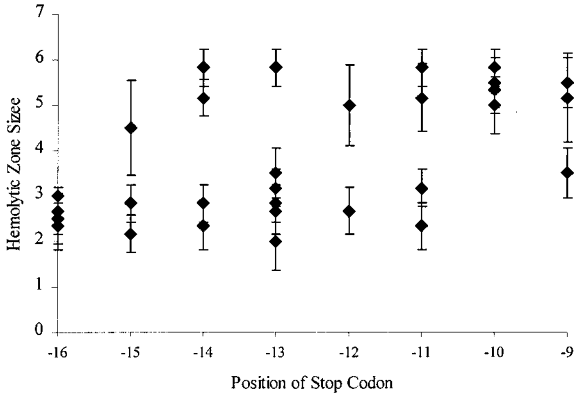

FIGURE 3: Effect of changing the length and amino acid composition of the C-terminus on secretion as demonstrated by Cterm1 stop mutants. The position of stop codon of 35 Cterm1 stop mutants is plotted against hemolytic zone size. Each data point represents a Cterm1 stop mutant, which can be viewed as a shorter version of Cterm2 random mutants (i.e., the amino acid composition of the extreme C-terminus is altered). The error bars are indicative of standard deviation (±1), calculated on the basis of six blinded measurements taken for each variant. Efficient transport is observed starting at position 2, and the chance of obtaining optimal secretion seems to be saturated after position 3. This suggests that a minimum of 13 residues is required after the linker region for efficient secretion. The bimodal distribution of Cterm1 stop mutants is also clearly illustrated here as most variants fall into either the top portion or the bottom portion.

Length Requirement for the Connector Region between Linker and Cterm2. A previous study (*6*) has shown that α2 random variants (Figure 1A) have a similar distribution to Cterm1 random variants (Figure 2A), suggesting that both the α2 and Cterm1 regions (11 + 8 = 19 residues within the signal sequence) could be replaced by any amino acids without having a major effect on transport! Since this stretch of amino acids is located between two important functional regions (Figure 1B), we refer to it as the “connector region”. If sequence specificity of the connector is not crucial for function, what role does this region play? To determine the relationship between the length of the connector region and transport efficiency, systematic internal deletions were cre-

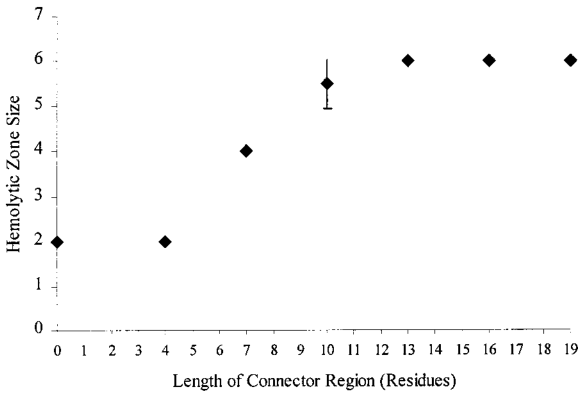

FIGURE 4: Effect of changing the length of the connector region on secretion efficiency. Systematic internal deletions of 3, 6, 9, 12, 15, and 19 residues were carried out in the connector region. All deletions started at position −27 (see Figure 1A). The error bars are indicative of standard deviation (±1), calculated on the basis of six blinded measurements taken for each variant. Consistent with results from the stop mutants (Figure 3), the plot above clearly illustrates that a minimum of 4−7 residues in the connector region is required for efficient secretion (i.e., 12−15 residues after the linker region).

ated in the connector and their hemolytic zone sizes were measured. A sigmoid curve was observed when the data were plotted (Figure 4). While removal of up to 12 residues would still allow efficient secretion, deletion of 15 residues or more would lead to a dramatic reduction in transport. On the basis of this experiment, we proposed that a minimum of 4−7 residues is required in the connector region for efficient secretion.

DISCUSSION

The *E. coli* hemolysin is a 107 kDa protein toxin that is transported directly from the cytoplasm to the outside of cells. Secretion is dependent on the hemolysin transporter complex spanning the inner and outer membranes, as well as a signal sequence located within the C-terminal 60 amino acids of hemolysin. Previous truncation studies have shown that the last 20 residues of the signal sequence are critical for transport. Remarkably, this region can tolerate a large number of point mutations, suggesting that the functional requirement(s) may not be at the primary sequence level. To investigate the functional specificity of the last 16 residues, we used a combinatorial approach to generate two contiguous 8 amino acid random libraries. It was found that no specific primary or secondary structural element is required between residues −16 and −9 since the Cterm1 region can be replaced by almost any combination of amino acids and still retain wild-type transport competence. However, correlation between the genotype and phenotype of the Cterm2 random variants provided evidence that hydrophobicity is a critical factor for efficient transport. In this context, hydrophobicity can be defined as the relative lack of positively charged residues, as well as the nonpolar nature of that region. Other than this apparent limitation, almost all combinations of amino acids in the Cterm2 region can be accommodated by the hemolysin transporter.

The combinatorial approach has also been used previously to investigate a conserved helix−linker−helix structural motif immediately upstream of the extreme C-terminus, leading to the elucidation of important structure−function relationships in that region (*6*). Integrating results from the previous and current study (six combinatorial libraries in total), we present a functional model of the hemolysin signal sequence that consists of two domains (Figure 1B). The first domain is 22 residues long and is comprised of the α1 helix as well as the linker region. An amphiphilic helical structure in the α1 region is supportive of efficient transport. The second domain covers the last 8 residues of the signal sequence, and a relative lack of positive charge appears to be the main determinant of transport efficiency. Connecting the two is a 19 amino acid connector region (i.e., the α2 and Cterm1 regions), with no specific sequence requirements.

It is intriguing that the connector region (which constitutes 19 out of 60 residues of the signal sequence) could be substituted by any primary sequence without having any major impact on transport. If the specific identity of amino acids in this region is not crucial, would the length be important? By carrying out systematic internal deletions of the connector region, we have shown that at least 4−7 residues are required between the two functional domains for efficient transport. This finding is further validated by the secretion pattern of α2 and Cterm1 stop mutants, in which efficient transport was observed only when 13 or more residues were present after the linker region. In other words, a minimum of 13 − 8 = 5 residues are required in the connector region (Figure 1B). In addition to demonstrating the length requirement, stop mutants have shown that a tail with few positively charged residues is a requirement for efficient secretion, providing additional evidence for the functional necessities at the extreme C-terminus.

The versatility of the hemolysin transporter system is illustrated by the remarkable collection of signal sequence variants that can be transported. However, since the transporter complex is not entirely nonselective, certain constraints must exist. The identification of subsets of sequences that support efficient transport has allowed us to deduce two principles of transport, providing an explanatory framework for the broad substrate specificity of the hemolysin system. First, we have demonstrated that the distinguishing features of the substrate are secondary structure (amphiphilic helix) and general biophysical property (relative noncharged nature/hydrophobicity), rather than primary sequence. Features such as hydrophobicity at the extreme C-terminus may facilitate insertion of the signal sequence into the inner membrane and/or the hydrophobic core of the transporter complex. Second, multiple features on the signal sequence contribute to transport in an incremental manner. Thus elimination of one feature is likely to decrease secretion efficiency but not abolish transport. This design could be seen as a preadaptation which allows the transporter to acquire novel substrates throughout evolution.

Principles of multiple substrate recognition highlighted in this study could also be applied to other members of the ABC transporter superfamily (*9*). For example, many compounds that are transported by the P-glycoprotein involved in multidrug resistance are relatively hydrophobic in nature, which is believed to facilitate their partitioning into the plasma membrane before interacting with the transporter (*10*). Likewise, the α1 amphiphilic helix and the relatively hydrophobic tail of the hemolysin signal sequence could be involved in a similar role. In a study parallel to ours, 100 substrates of P-glycoprotein were analyzed for features in common (*11*). While many of these substrates have distinct structure, it was found that they all contain multiple electron-donor groups with a fixed spatial separation. These groups are believed to function as recognition elements for substrate binding to the transporter, not unlike the functional domains within the hemolysin signal sequence. The functional model proposed in this study provides a platform for further experimentation to elucidate the precise nature of the interaction between the two domains of the hemolysin signal sequence and transporter complex. This is likely going to provide further insights into the mechanistic aspects of a transport process shared by many ABC proteins.

While combinatorial analysis of the hemolysin signal sequence has been helpful in identifying features that determine substrate specificity, this approach can be applied in a more general context to investigate other biological systems, such as the N-terminal membrane targeting signal. Indeed, this technique is not limited to the study of protein function. It can be equally effective in analyzing the sequence requirement of specific nucleic acid regions, including transcription activator domains. The large population of variants generated provides the investigator with a rich data set to determine if a restricted subset of sequences is required for function in a particular region and allows for a greater degree of confidence in the interpretation of results. As demonstrated in this study, random libraries can be quite informative in exploring the sequence limitation within a selected biological window. Still, the more powerful application of combinatorial analysis is in answering highly refined questions with a more specific library design.

ACKNOWLEDGMENT

We are grateful to Dr. Jaclyn Hung, Isaac Tan, and Steven Ralph for assistance in this study. We acknowledge Dr. Fang Zhang and Carla Morden for initiating some of the earlier work in utilizing a combinatorial approach to investigate the hemolysin system. We thank Dr. Nhu Le for assistance in statistical analysis, Dr. Eric Cabot for assistance in programming, and Drs. Mark Blight and Barry Holland for generously providing the anti-hemolysin antiserum, as well as Dr. Jonathan Sheps and Barbara Schmidt for critical review of the manuscript.

SUPPORTING INFORMATION AVAILABLE

Two tables listing the genotype and phenotype of all combinatorial mutants, provided for reference. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Mackman, N., Baker, K., Gray, L., Haigh, R., Nicaud, J. M., and Holland, I. B. (1987) *EMBO J. 6*, 2835−2841.
2. Chervaux, C., and Holland, I. B. (1996) *J. Bacteriol. 178*, 1232−1236.
3. Koronakis, V., Koronakis, E., and Hughes, C. (1989) *EMBO J. 8*, 595−605.
4. Zhang, F., Greig, D. I., and Ling, V. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 4211−4215.
5. Higgins, C. F. (1992) *Annu. Rev. Cell Biol. 8*, 67−113.
6. Hui, D., Morden, C., Zhang, F., and Ling, V. (2000) *J. Biol. Chem. 275*, 2713−2720.
7. Ludwig, A., Vogel, M., and Goebel, W. (1987) *Mol. Gen. Genet. 206*, 238−245.
8. Stanley, P., Koronakis, V., and Hughes, C. (1991) *Mol. Microbiol. 5*, 2391−2403.
9. Zhang, F., Sheps, J. A., and Ling, V. (1993) *J. Biol. Chem. 268*, 19889−19895.
10. Shapiro, A. B., and Ling, V. (1998) *Eur. J. Biochem. 254*, 181−188.
11. Seelig, A. (1998) *Int. J. Clin. Pharmacol. Ther. 36*, 50−54.

BI011425G