I. Preparation of the files. You need two files: a nucleotide fasta and a fasta file with all the translated peptides.

IA. The nucleotide fasta file. Save the fasta file with the header ">fosmidname [organism=uncultured bacterium]".

IB. The amino acid fasta file. This you generate from a genbank file that has the translations and annotations. It is easiest if you name the genbank file with the fosmid name, as that will be used as the prefix for the ORF numbers.

**this script does not do any predictions! It only sorts the data in the .gbk file. Make sure all of the stuff in that file is what you want to submit to Genbank!!

The gbk_aa_sequin.pl script takes a genbank file with the name fosmidname.gbk and produces a fasta file of the predicted translated proteins, each with a header line that Sequin can understand. Each protein gets assigned a number x or xc, depending on whether it is on the forward or reverse strand.

The header lines will be ">fosmidname_number [gene=ORF_number] [protein=annotation]". These lines and the translations are what Sequin needs. The two numbers are the same.

To use:
-copy the script to whatever folder you like.
-Type on the command line "perl scriptfolder/gbk_aa_sequin.pl folder/fosmidname.gbk"
-The output file will be fosmidname.gbk.aa.

Before moving on, check the .aa file. I think I fixed the bugs this script had, but I cannot be sure.

II. Sequin.
A. Information about the paper. Fill in the blanks on the first four screens. When you get to the screen where you choose what kind of sequence it is, click on the File menu and then choose Export submission information. Save that stuff so you don't have to type the same information for every fosmid. This only matters if you are submitting >1 fosmid sequence for the same paper.

B. Import the nucleotide fasta file. Under the "Organisms, Locations, and Genetic Codes" tab, make sure "Genetic Code" is set to Bacterial and Plant Plastid. FOR ENVIRONMENTAL FOSMID CLONES: Under "Source Modifiers", choose "Metagenome", "Environmental Sample," and "Source Location." Type in something on the source location line. "Lineage" should be set to Bacteria.

C. Under the "Proteins" tab, import the .aa fasta file. Say ok to everything. I skipped the "annotations" tab altogether, i didn't find anything helpful there. Go to the next form and then click "Done."

D. A list of errors will come up. The ones that are related to partial translations and protein titles can safely be ignored. The issues that I had with Genbank subsequent to submitting sequences prepared this way only had to do with their unwillingness to believe in GTG start codons, and nothing to do with anything else. Go ahead and submit the file. Good luck!

Hope this works for you.... Cheers.