SEQUENCE COMPLIANCE SOUP TO NUTS

Bob Wax
Quality Assurance Specialist
Technology Center 1600

Why do we have the sequence rules?

- Search
  - Automated Biotechnology Sequence Search (ABSS) System
  - Prior art databases searched
    - Protein: A_Geneseq, UniProt, PIR and Published_Applications_AA, Issued_Patents_AA
    - Nucleic: N_Geneseq, GenEmbl, EST and Published_Applications_NA, Issued_Patents_NA

Why do we have the sequence rules?

- Search
  - Interference databases searched
    - Easy system for examiners to use to detect potentially interfering sequence subject matter
    - Results accessible only to examiners

Why do we have the sequence rules?

- Publication
  - National Center for Biotechnology Information (NCBI)
    - The USPTO exports patented and published sequence listings to NCBI in GenBank's format (asn) so they can more easily be published
  - Publication Site for Issued and Published Sequences (PSIPS)
    - Sequence listings at least 300 pages (roughly 600Kb) are published at this USPTO website

What are these rules anyway?

- US Rules – 37 CFR 1.821-825
  - Original rules: Effective October 1, 1990 (see Federal Register, Vol. 55, No. 84, May 1, 1990, p. 18230)
  - Amended rules: Effective July 1, 1998 (see Federal Register 63:104, 29620-29643, June 1, 1998)

What are these rules anyway?

- International Rules - WIPO Standard ST.25, effective July 1, 1998
  - http://www.wipo.int/scit/en/standards/pdf/st_25.pdf

How do I comply with the sequence rules?

- Manually type the sequence listing while referring to the sequence rules
  - Not recommended – time consuming, error prone
- Use software such as PatentIn
  - Free software provided by USPTO
- Other software
  - FastSeq

Common compliance pitfalls

- The inclusion of sequences containing fewer than four (4) specifically defined amino acids or ten (10) nucleotides (four specifically defined) is not mandatory (37 CFR 1.821(a))
  - Unless there is some important reason for including them, their submission is discouraged

Common compliance pitfalls

- Examples of specifically defined amino acids
  - Ile, Pro, Glu
  - Xaa, defined as Pro
- Examples of non-specifically defined amino acids
  - Xaa, defined as (for example)
    - any of Ile, Pro and Glu
    - any naturally-occurring amino acid

Common compliance pitfalls

- The organism of each sequence must be defined at heading <213> (Organism) (37 CFR 1.822(b))
- Genus/species or “artificial sequence” or “unknown”
  - If artificial sequence or unknown, further definition is required at headings <220> - <223>
  - Use Genus/species if at all possible
    - If it is a human sequence, for example, use Homo sapiens
    - Depends on source of the actual sequence
      - Does not matter if isolated or synthesized

Common compliance pitfalls

- Artificial Sequence
  - Explain why you consider the sequence artificial
    - Sequence per se is derived from human thought
    - Several sequences piece together
      - Use Synthetic construct

Common compliance pitfalls

- Unknown
- Use if there is no scientific name disclosed or only a partial scientific name, e.g., Bacillus sp.
- Use if only the source of the organism is disclosed, e.g., “soil sample from Pittsburgh”
  - Example sequence listing section:
    - <213> Unknown
    - <220>
    - <223> Bacillus species

Common compliance pitfalls

- The specific location of each variable (“n” or “Xaa”) in a sequence must be identified and explained at each specific location in the sequence (37 CFR 1.822(b))
  - PatentIn can do this automatically
- “n” and “Xaa” may only be used to represent a single nucleotide or amino acid, respectively, and may not be used to represent a label or reporter molecule or some other moiety
  - Such moieties should not appear in the sequence listing

Common compliance pitfalls

- For a variable-length string, present the largest embodiment of the sequence and the specific variables, including absent bases/residues, in fields <220> - <223>, also called the feature section
- For example, the sequence Ile Pro Xaa_6 Glu Asp would be shown as:
  - <220>
  - <221> MISC_FEATURE
  - <222> (3)..(8)
  - <223> Xaa at positions 3-8 may be any naturally-occurring amino acid and up to five of them may be absent
  - <400> 1
  - Ile Pro Xaa Xaa Xaa Xaa Xaa Xaa Glu Asp

Common compliance pitfalls

- Nucleotide sequences must be presented as single stranded, oriented 5' to 3', left to right (37 CFR 1.822(c)(5))
  - For double stranded DNA show only the sense strand
    - If the invention lies in the antisense strand, also provide that as a separate sequence, identified as the antisense strand of the complementary sequence
    - May need to use a sequence manipulation tool to display antisense sequence 5’ to 3’

Common compliance pitfalls

- Amino acid sequences must be presented oriented as amino to carboxy, left to right (37 CFR 1.822(d)(3))
  - Leave off the $_2$HN- and –COOH groups

Common compliance pitfalls

- Amino acid sequences containing even one D-amino acid are excluded from the sequence rules (37 CFR 1.821(a)(2))
- However, voluntary submission of these sequences is encouraged to aid in searching
    - Such sequences could be submitted with the corresponding L-amino acid with a feature defining it as D

Common compliance pitfalls

- The computer readable form (CRF) of the sequence listing must be filed as ASCII text only (with extension .txt or .app) (37 CFR 1.824(a)(2))
- CRFs that are submitted as a word processing file (e.g., having extensions such as .doc or .wpd) or as a PatentIn project file (with extension .prj) will not be accepted

Common compliance pitfalls

- Publicly known sequences included in an application for any purpose must be included in the Sequence Listing (37 CFR 1.821(c))
  - Rule of Thumb
    - If a sequence is disclosed it must be included in the sequence listing

Common compliance pitfalls

- Fragments of larger sequences do not need to appear in the Sequence Listing as long as they are identified in the application as specific portions of a larger sequence, which is included in the formal Sequence Listing (e.g., residues 1-25 of SEQ ID NO: 15)
  - Inclusion of such fragment sequences in the Sequence Listing as their own identification number is permitted but discouraged

Common compliance pitfalls

- Sequences having a gap or gaps must be displayed as separate sequences in the Sequence Listing. For example, if a chemical moiety has several strands of protein attached to it, each protein sequence should appear in the Sequence Listing separately. The chemical moiety should NOT be shown (37 CFR 1.822(e))
- Sequences made of fragments of other sequences must be displayed as separate sequences in the Sequence Listing (37 CFR 1.822(e))

Common compliance pitfalls

- Sequence Listings often lack compliance because of minor formatting issues
- Use of PatentIn minimizes such occurrences
  - Occasionally, PatentIn's "Copy to Disk" function results in loss of hard returns on the CRF
    - to correct, regenerate the Sequence Listing and use Windows Explorer to copy the text file to the CRF

Common compliance pitfalls

- Improper CRF transfer requests
  - Proper request includes
    - Request to transfer the CRF
    - Paper copy of sequence listing (not transferable)
    - Statement that they are the same
    - Statement that there is no new matter
    - See (37 CFR 1.821(c))

Common compliance pitfalls

- Improper CRF transfer requests

  - failure to include the statements that need to be present (CRF and paper copy identical; no new matter)

  - failure to include sequence listing in PDF form when requesting transfer via EFS

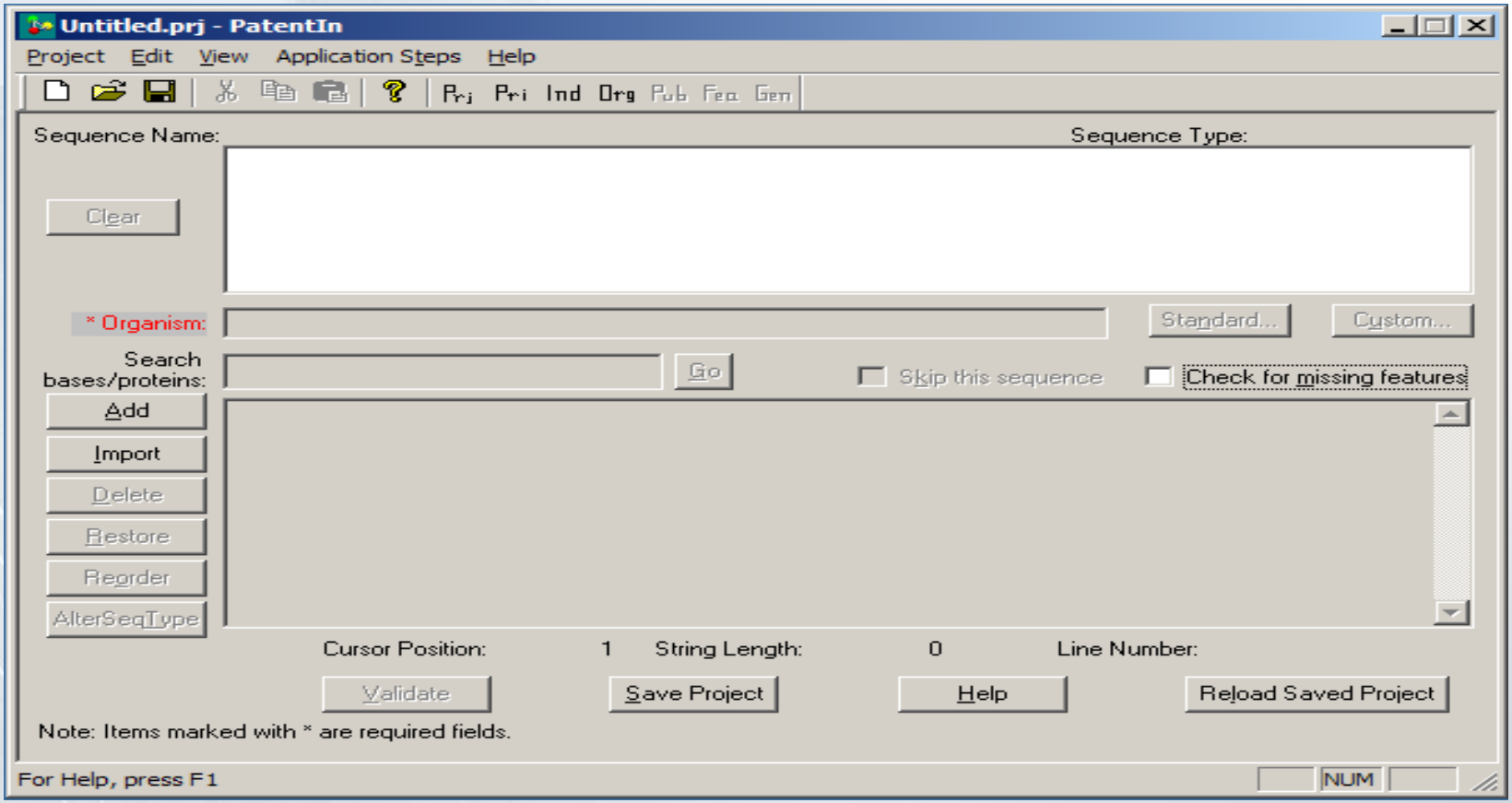
  - mistakenly filing both a CRF transfer request and an ASCII sequence listing when only one is needed

PatentIn

- What is it?
  - Sequence listing authoring software provided by the USPTO
- Where do I get it?
  - http://www.uspto.gov/web/offices/pac/patin/patentinrel.htm
- How do I use it?
  - User manual can be found at the above link

PatentIn

Screen Shot of PatentIn 3.5

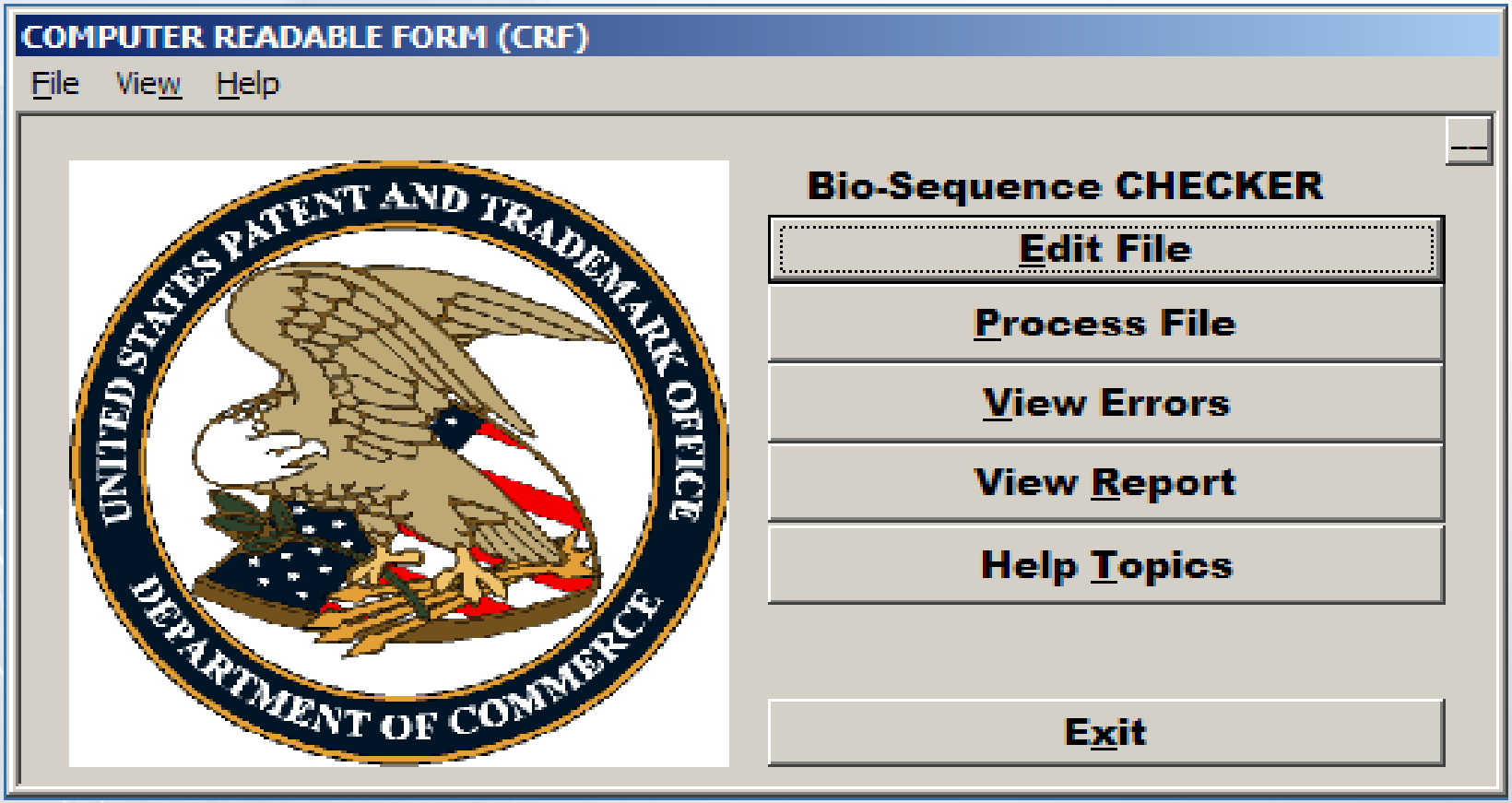


Checker

- What is it?
  - Verification software provided by the USPTO for preliminary evaluation of sequence rule compliance
- Where do I get it?
  - http://www.uspto.gov/web/offices/pac/checker/

Checker

Screen Shot of Checker



Checker

- How do I use it?
  - User manual can be found at the above link
- Warning
  - Checker DOES NOT validate whether information in free text fields is proper

Checker

- Common problem
  - Checker sometimes gives you the message, “Input file is neither numeric nor alpha”
  - This is almost always caused when an inventor’s name has a non-English symbol such as an e with an accent over it
    - Fix by changing the letter to an equivalent English symbol, e.g., an e without the accent, run Checker again, then put the original letter back before submitting the sequence listing

Filing options

- Diskette (or CD) and paper
  - (37 CFR 1.824)
- 3 CDs
- Electronic Filing System (EFS)
  - Legal Framework (http://www.uspto.gov/ebc/portal/efs/legal.htm)

Filing options

- Diskette (or CD) and paper
  - Copy sequence listing onto a floppy disk or CD (thus creating the CRF)
  - Print the sequence listing on paper
    - include a statement that the CRF and the paper copy are the same
      - if filing in response to a Notice to Comply also include a statement the there is no new matter

Filing options

- 3 CDs
  - Copy sequence listing onto a CD-ROM
  - Can use CD-R if the disk is finalized after recording the CRF, but NOT CD-RW
  - Make two copies
  - Label one as the CRF (see 37 CFR 1.824(a)(6), label the second as Copy 1 and label the third as Copy 2

Filing options

- Electronic Filing System (EFS)
  - Learn about EFS at this website: http://www.uspto.gov/ebc/efs_help.html
    - Add the sequence listing to your EFS-Web submission
    - No paper copy or statement needed for initial filing
      - If filing in response to a Notice to Comply a statement that there is no new matter is needed.
      - Sequence listing is automatically processed by SCORE and immediately placed in ABSS (if compliant)

Notice to Comply

- Who sends them?
  - The Office of Patent Application Processing (OPAP)
- Time period to respond
  - Two months, extendable to six months under 37 CFR 1.136(a) or (b)

Notice to Comply

- Where to get help
  - Call the person in OPAP who signed the Notice to Comply
  - Call Mark Spencer (STIC Systems Branch) at (571) 272-2533
  - Call Bob Wax (QAS, TC 1600) at (571) 272-0623 for particularly thorny questions involving sequence rule interpretation

Common errors in sequence listings

- Following are some common errors found during verification of the sequence listing
  - You will see some of these on your Notice to Comply with the Sequence Rules

Common errors in sequence listings

- Numeric identifier <213> is something other than “Scientific name, i.e., Genus/species, Unknown or Artificial Sequence”
  - Fix by changing answer in field <213>
- Insufficient or missing explanation in numeric identifier <223> for “<213> Artificial Sequence” or “<213> Unknown”
  - Fix by providing better explanation

Common errors in sequence listings

- Amino acid designators not starting with a capital letter
- Sequence listing not in English language
- Sequence listing not in ASCII text format
  - Fix for these obvious

Common errors in sequence listings

- Missing or incorrect information in mandatory feature for use of “n” or “Xaa” in the sequence
  - If n or Xaa appears it MUST be further defined
    - Fix is to provide the definition
- Extra text or symbol at the end of the file, after the last sequence
  - Fix is to delete the text

Common errors in sequence listings

- Numeric identifier <160> (number of sequences) does not match the number of sequences in the file
- Numeric identifier <211> (length) does not match the total number of residues in the sequence
  - Fix is to correct the information
- Sequence listing is not in valid format, per Sequence Rules
  - Simple listing of sequences rather than a "Sequence Listing", e.g. SEQ ID NO: 1, followed by the sequence, etc.
  - Partial sequence listings, e.g., the application info header (<110> to <170>) is absent

Common errors in sequence listings

- Missing field <130> (File Reference)
  - Required for every sequence listing
    - Usually attorney docket number
- Missing fields<140> (Current Application Number) and <141> (Current filing date) when required
  - Not needed for new filing
  - Needed for filing corrected sequence listing
    - Usually when replying to a Notice to Comply

FAQs

- Do genes identified by gene accession numbers in the specification need to comply with the sequence rule requirements?

FAQs

- No, they are not considered disclosures of sequences
- When accession numbers appear in claims, however, they may raise an issue of improper incorporation of essential material by reference
  - If the sequences need to be brought into the disclosure then they must comply with the sequence rules

FAQs

- Do sequence rules apply to reissue and continuation applications?
  - Absolutely. The CRF does not carry over from the parent file so sequence compliance must be perfected again
  - You can do this by requesting transfer of the CRF or by filing a new copy of the sequence listing

FAQs

- How do I comply if my application discloses a repeat of sequences, some of which are identical and some of which are not?
  - Three categories of repeat
    - Repeats of bases within a sequence
    - Repeated disclosures of the same sequence in the specification
    - Large pyramid of overlapping sequences where each new sequence just adds some bases to the sequence before

FAQs

Repeats of bases within a sequence

- $(cgatgccaatt)_4$
  - Enter either of two ways
    - cgatgccaatt with explanation that it is repeated 4 times
    - gatgccaattcgatgccaattcgatgccaattcgatgccaatt

FAQs
Repeats of bases within a sequence

- $(atgg)_n(cggc)_m$
  - If n=2-4 and m=3-5, for example, put in the largest number of repeats and add a feature saying some of them may be absent:

  - <220>
  - <221> misc_feature
  - <222> (1)..(6)
  - <223> these nucleotides may be absent

  - <220>
  - <221> misc_feature
  - <222> (17)..(23)
  - <223> these nucleotides may be absent

  - <400> 1
  - atggatggatggatggcggccggccggccggccggc

FAQs

Repeats of bases within a sequence

- $(atgg)_n$ and $(cggc)_m$ listed separately
  - No compliance necessary
  - Have the claim recite atgg repeated n times is joined (via a phosphodiester bond) 3' to 5' to cggc m times

FAQs

Repeats of bases within a sequence

- Pro Glu Arg Asp Xaa_n Ile Tyr His Cys
  - Where n must be a positive integer
  - List as two sequences separated by an undefined group, treating the infinitely repeated Xaa as a chemical moiety

    <400> 1

    Pro Glu Arg Asp

    <400> 2

    Ile Tyr His Cys

FAQs
Repeats of bases within a sequence

- Leu Arg Xaa_{3-6} Cys Tyr
  - List the largest number of repeats and add a feature saying some of them may be absent
    - Leu Arg Xaa Xaa Xaa Xaa Xaa Xaa Cys Tyr
    - Feature: amino acids at positions 3-5 may be absent

      <220>

      <221> MISC_FEATURE

      <222> (3)..(5)

      <223> these amino acids may be absent

      <400> 1

      Leu Arg Xaa Xaa Xaa Xaa Xaa Xaa Cys Tyr

FAQs

- Repeated disclosures of the same sequence in the specification
  - put the same SEQ ID # next to each repeat
    - Do not assign a new SEQ ID # to each repeated sequence
- Large pyramid of overlapping sequences where each new sequence just adds some bases to the sequence before
  - Provide a SEQ ID # for the largest sequence in the series and identify the rest as locations within the larger sequence

FAQs

- What is the definition of a branched amino acid sequence?
  - A branched amino acid sequence is one where one or more amino acids branch off the main chain via a peptide bond to an amine group on an amino acid side chain, e.g., Lysine (**H_2N-**$CH_2CH_2CH_2CH_2CH(NH_2)COOH$**)**

FAQs

- Disulfide bonds DO NOT create a branched sequence
  - Interchain disulfide bond between two seqeunces
  - Intrachain disulfide bond within a single sequence

FAQs

- Would electronic filing get the sequences approved and entered properly into the database as opposed to paper filing?

FAQs

- Either way you file the sequence listing will get entered correctly if it is in compliance
- EFS is much easier for the applicant and is automated at the USPTO
  - lack of human involvement permits entry of compliant sequence listings faster than before the automated system was implemented

FAQs

- I had my sequence listing prepared via PatentIn. Why did my sequence listing submission still get rejected by the patent office?
  - The internal verification software the USPTO uses to verify sequence listings is called CRF
  - Checker is similar to CRF but not identical
    - Information provided in field <223> for artificial sequence or unknown organism must be manually verified

FAQs

A major reason for noncompliance is that the information provided in field <223> to explain an artificial or unknown organism is improper

- Indicating what the artificial sequences are is acceptable, e.g., primer, aptamer, linker, adapter, cloning vector, expression vector, siRNA, probe, expressed sequence tag, etc.
- Chimeric constructs should identify sources of the parts, etc.

FAQs

- Should I leave field <140> (Current Application Number) empty when there is no assigned serial number? Should I wait to file my sequence listing until a serial number is assigned?
  - If you are filing a sequence listing for a new case there is no assigned serial number.
  - Don’t wait to file until a serial number is assigned
  - Leave fields <140> and <141> (current filing date) empty
  - Remember that field <130> (File reference) is required

Q and A

Any Questions?

Bob Wax
Quality Assurance Specialist
Technology Center 1600
571-272-0623
robert.wax@uspto.gov