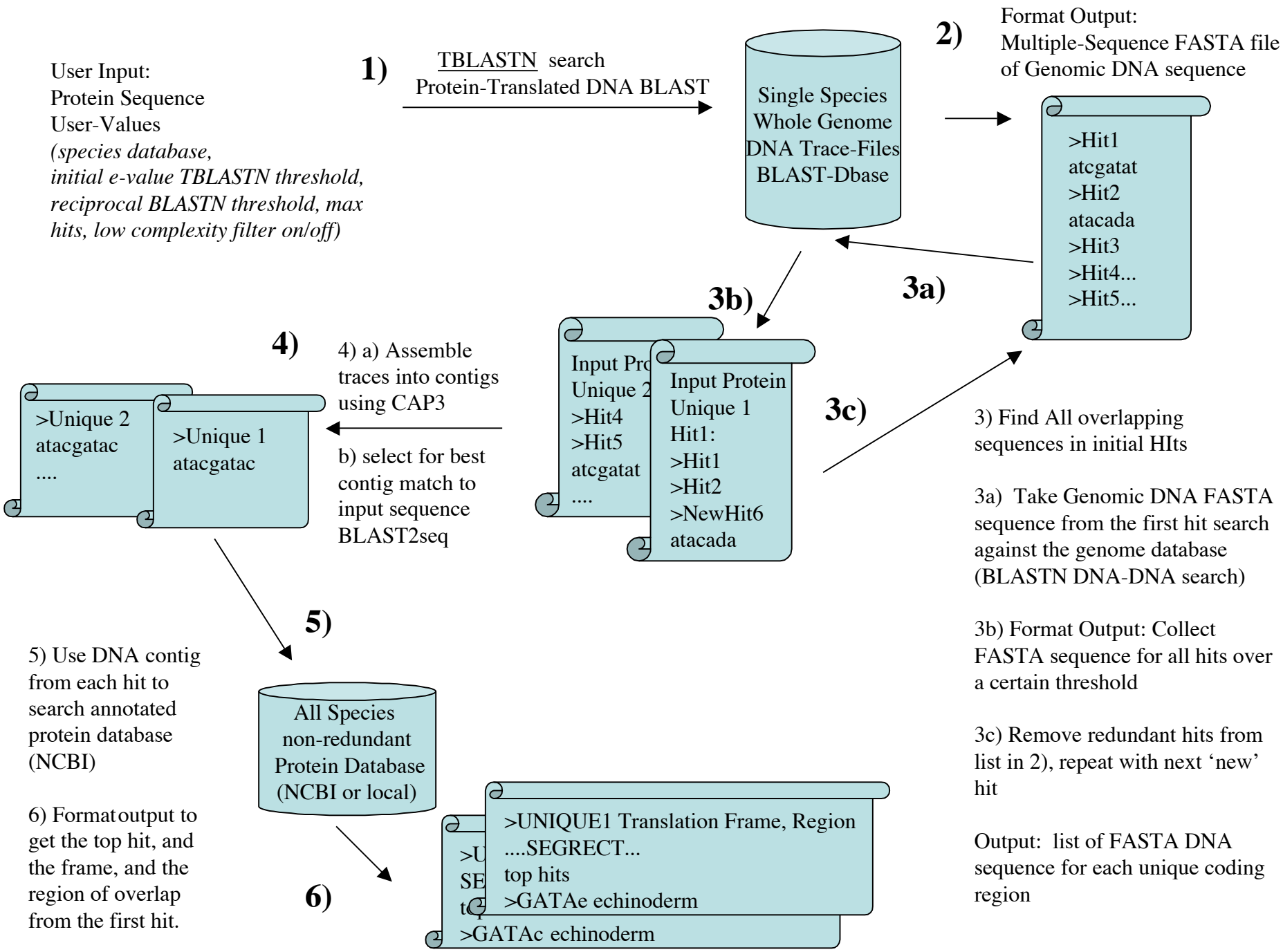
User Input:
Protein Sequence
User-Values
(species database,
initial e-value TBLASTN threshold,
reciprocal BLASTN threshold, max
hits, low complexity filter on/off)
1)
TBLASTN search
Protein-Translated DNA BLAST
Single Species
Whole Genome
DNA Trace-Files
BLAST-Dbase
2)
Format Output:
Multiple-Sequence FASTA file
of Genomic DNA sequence
>Hit1
atcgatat
>Hit2
atacada
>Hit3
>Hit4...
>Hit5...
3a)
3b)
Input Pro
Unique 2
>Hit4
>Hit5
atcgatat
....
Input Protein
Unique 1
Hit1:
>Hit1
>Hit2
>NewHit6
atacada
3c)
3) Find All overlapping
sequences in initial HIts
3a) Take Genomic DNA FASTA
sequence from the first hit search
against the genome database
(BLASTN DNA-DNA search)
3b) Format Output: Collect
FASTA sequence for all hits over
a certain threshold
3c) Remove redundant hits from
list in 2), repeat with next 'new'
hit
Output: list of FASTA DNA
sequence for each unique coding
region
4)
4) a) Assemble
traces into contigs
using CAP3
b) select for best
contig match to
input sequence
BLAST2seq
>Unique 2
atacgatac
....
>Unique 1
atacgatac
5)
5) Use DNA contig
from each hit to
search annotated
protein database
(NCBI)
6) Format output to
get the top hit, and
the frame, and the
region of overlap
from the first hit.
All Species
non-redundant
Protein Database
(NCBI or local)
6)
>UNIQUE1 Translation Frame, Region
....SEGRECT...
top hits
>GATAe echinoderm
>GATAc echinoderm