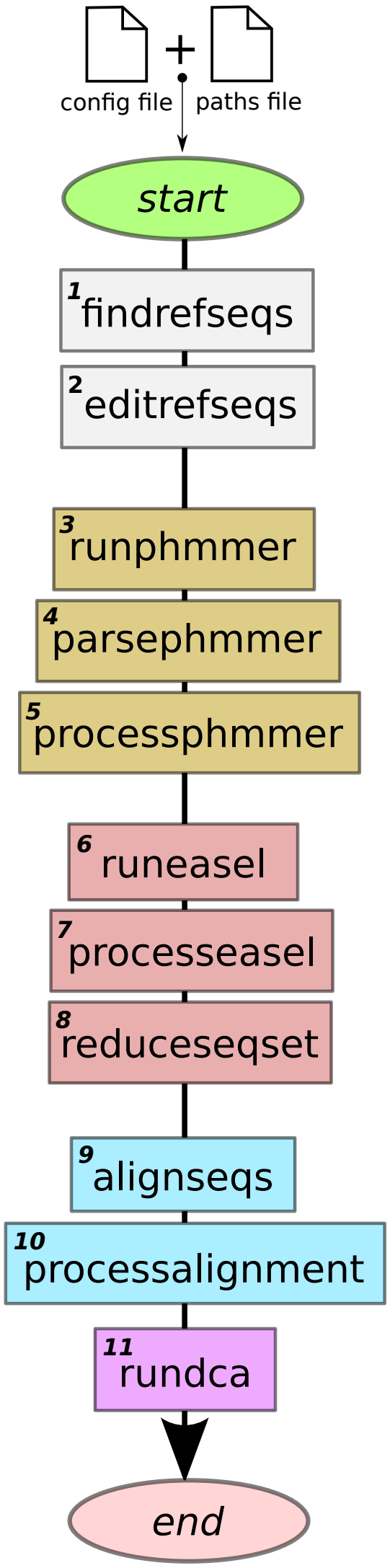

run_workflow.py

	file formats	***descriptions of steps***
1 findrefseqs		Finds paths to reference fasta sequence files.
2 editrefseqs	*xxxx_A_refseq.fasta* *xxxx_B_refseq.fasta*	Edits refseqs by removing nonstandard residues.
3 runphmmer	*xxxx_A_refseq_phmmer.log* *xxxx_B_refseq_phmmer.log*	Runs PHMMER search on both reference seq fastas against UniProt (release 4.21)
4 parsephmmer	*xxxx_A_refseq_phmmer.keyfile* *xxxx_B_refseq_phmmer.keyfile*	Parses accession IDs from both phmmer logfiles into keyfiles
5 processphmmer	*xxxx_A_refseq_phmmer_matched.keyfile* *xxxx_B_refseq_phmmer_matched.keyfile*	Checks for min/max nr of accIDs per keyfile, matches accIDs based on organism
6 runeasel		Runs HMMER easel to extract sequences from Uniprot based on accid.
7 processeasel	*xxxx_A_refseq_phmmer_matched.fasta* *xxxx_B_refseq_phmmer_matched.fasta*	Removes seqs from both fasta files corresponding to organisms that could not be extracted (if any).
8 reduceseqset		Removes seqs from both fasta files that are above a certain length (default: 1600). Does this to avoid Muscle aln memory errors.
9 alignseqs	*xxxx_A_refseq_phmmer_matched.aln* *xxxx_B_refseq_phmmer_matched.aln*	Aligns sequences in both fasta files
10 processalignment	*Joint_xxxx_A_xxxx_B_aln.fasta*	Joins matched sequences together to make a joint alignment file
11 rundca	*Joint_xxxx_A_xxxx_B_aln_mfdca_scores .dat,etc.*	Runs DCA on joint alignment, gives out a scores file