

## Identyfikacja niekodujących RNA

### Identyfikacja genów z użyciem modelu kowariancji [Infernal]

1. Wejdź na stronę RFAM (<http://rfam.xfam.org>) i wyszukaj rodzinę ncRNA: PrfA.
2. Pobierz alignment dla członków tej rodziny w formacie Stockholm
3. Obejrzyj alignment bezpośrednio na stronie z użyciem wtyczki **jalview**. Obejrzyj ten sam alignment pobrany w formacie Stockholm w edytorze tekstu. Zapisz do protokołu jaka dodatkowa informacja jest zawarta w formacie Stockholm? Opis formatu znajdziesz tutaj: [https://en.wikipedia.org/wiki/Stockholm\\_format](https://en.wikipedia.org/wiki/Stockholm_format)
4. Zbuduj model kowariancji używając komendy **cmbuild** z pakietu **infernal**. Najpierw przejrzyj dostępne opcje i następnie uruchom program.
5. Przeanalizuj otrzymany raport oraz uzyskany model kowariancji w edytorze tekstu. Na podstawie entropii obliczonej dla modelu kowariancji (CM) oraz modelu markowa (HMM) zapisz do protokołu zysk w wyniku zastosowania informacji zawartej w modelu kowariancji.
6. Uruchom kalibrację uzyskanego modelu za pomocą komendy **cmcalibrate**. Najpierw przejrzyj dostępne opcje. Zwróć uwagę na opcje **--forecast**.
7. Przeszukaj sekwencję bakteryjną z *Listeria monocytogenes* (pobierz w formacie fasta z <http://www.ebi.ac.uk/ena/data/view/AY512499>) za pomocą skalibrowanego modelu w celu identyfikacji nowych członków rodziny PrfA za pomocą komendy **cmsearch**. Zwróć uwagę na opcję **--tabfile**. Zapisz do protokołu ile trafień udało się uzyskać.
8. Wybierz sekwencje o score'em większym od 100 za pomocą **esl-sfetch**
9. Przeprowadź alignment uzyskanych sekwencji do modelu za pomocą komendy **cmalign**. Sprawdź dopasowanie zidentyfikowanych sekwencji do modelu. Zapisz do protokołu, która z sekwencji najlepiej spełnia założenia modelu strukturalnego? Czy miała ona najwyższą punktację?

## Identyfikacja niekodujących RNA w genomie *Saccharomyces cerevisiae* za pomocą RNAz

1. Pobierz dane: <https://www.tbi.univie.ac.at/papers/SUPPLEMENTS/MiMB/yeast-examples.tar.gz>
2. Pliki potrzebne do analizy:
  - Input.maf - dopasowanie sekwencji międzygenowych pochodzących z 7 gatunków *Saccharomyces*
  - sgdRNA.bed - adnotacja genomowa
3. Preprocessing dopasowania sekwencji (input.maf) za pomocą rnazWindow.pl.
  - a. Podziel na alignment na nachodzące okna wielkości 120 z przesunięciem o 40.
  - b. Usuń sekwencje zawierające więcej niż 25% przerw.
  - c. Usuń sekwencje o długości < 50 nt i zawartości GC > 0.75
  - d. Ustaw minimalną liczbę sekwencji na 4
4. Uruchom program RNAz na pliku z wygenerowanymi oknami z poprzedniego kroku. Ustaw dopuszczalną wartość P-value na 0.5. Oblicz wynik dla obu nici. Zastosuj parametr `—no-shuffle`.
5. Sklastruj nachodzące się okna w pojedyncze loci za pomocą rnazCluster.pl. Zapisz do protokołu liczbę okien przed klastrowaniem oraz uzyskaną liczbę loci.
6. Przefiltruj uzyskane wyniki, tak aby pozostały te o  $P > 9$  ( rnazFilter.pl ) oraz posortuj wg. combPerPair (rnazSort.pl)
7. Zwizualizuj wyniki za pomocą rnazIndex.pl `—html`
8. Przeprowadź adnotację uzyskanych wyników za pomocą (rnazAnnotate.pl)
9. Ile ze zidentyfikowanych loci zostało zaadnotowanych?
10. Jakie rodziny ncRNA zostały zidentyfikowane?