PROTEIN SEQUENCING WITH AN ADAPTATIVE GENETIC ALGORITHM FROM TANDEM MASS SPECTROMETRY

Jean-Charles BOISSON², Laetitia JOURDAN¹, El-Ghazali TALBI¹ and Christian ROLANDO²

¹LIFL/INRIA Futurs, UMR USTL/CNRS 8022 ²LCOM, UMR CNRS 8009
Lille, FRANCE

RESULTS

ADAPTIVITY

EVALUATION

IDENTIFY:
- know protein variants,
- unknown proteins.

MANNER:
- Find the protein sequence from experimental data without databases,
- inspired from classic de novo peptide sequencing method,
- exploitation of two types of data:
  - MS spectra (mass / charge) / intensity spectrum corresponding to the digestion of one protein or a protein mix into peptides.
  - MS/MS spectra (mass / charge) / intensity spectrum corresponding to the fragmentation of one peptide into ions.

METHOD:
- First step is exploitation of an experimental MS spectrum to find the peptide chemical formula of the experimental protein:
- Design of an adaptive Genetic Algorithm (GA).

CODING: an individual is a peptide list, a peptide is an array of amino acids that can have modifications.

INDIVIDUAL PEPTIDE LIST \rightarrow \text{TRANSFORMATION} \rightarrow \text{ISOTOPIC DISTRIBUTION COMPUTATION} \rightarrow \text{COMPARISON WITH EXPERIMENTAL SPECTRUM} \rightarrow \text{GLOBAL FITNESS COMPUTATION} \rightarrow \text{INDIVIDUAL FITNESS}

Figure 1

Evaluation steps to make for each peptide of the individual

ADAPTIVE MUTATION UTILITY:
- difficulty to find the right probability of each mutation (6 in our case),
- impact of each mutation on the individual fitness of the population evolution.

ADAPTATIVE MUTATION IMPLEMENTATION:
- when a mutation is chosen, its application can:
  - improve the individual fitness (case 1),
  - let the individual fitness unchanged (case 2),
  - penalize the individual fitness (case 3).
- in order to compute the new mutation rate, a progress factor is calculated according to the impact of the mutation on the individual fitness (case 1, 2 or 3).

ADAPTIVE MUTATION RESULTS:
- Figure 1 and 2 show the mutation rate evolution on two runs with the same configuration (without post-translational mutation).
- In the two cases, the mutation rates need more than 500 generations to be stable.
- According to the initial population, the mutation rate evolution is very different before stabilization.

Figure 2

RESULTS

DATA USED:
- experimental and theoretical MS spectra:
  - experimental data are provided by the LCOM,
  - in the theoretical spectrum generation, the intensities are not realistically computed (just normalized),
- 2 main studied proteins:
  - human Apolipoprotein-AI (Apo-AI),
  - bovine cytochrome-C (Cytc-C).

RESULTS:
- Figure 3 is compound of two spectra: the Apo-AI simulated spectrum given as the MS Experimental spectrum to the GA and the best individual spectrum after the evolution.
- Figure 4 is a zoom on one peak of the figure 3.
- The most of the peptide chemical formulas are found and the best individual have a peptide mass fingerprinting that corresponds well to the experimental one.
- Having the right peptide chemical formulas do not correspond often to the right peptide sequences and do not give information about the peptide order in the experimental one.
- The first step of a new method for protein variant and unknown protein identification has implemented and tested,
- the chosen optimization method (the genetic algorithm) and the use of adaptive mutations give good results,
- this step can be improved in order to get exactly the right number of different peaks.
- our approach is able not to differentiate two or more peptides with the same mass.

Figure 3

Figure 4

PERSPECTIVES

- improving the theoretical spectrum generation \( \odot \) real intensity computation,
- designing the two next steps of our approach \( \square \) first version already done.
- testing the different steps of our approach on other proteins with higher length and mass \( \square \) a parallelized version of our GA already exist but without adaptive mutation (acknowledgement to A8 GRID for the tests made on the French grid GRID5000),
- validating the complete approach on identified protein variant.