

Lectures de tesis doctorals

Anunci del darrer pas cap al doctorat: la lectura.



[\(English Version\)](#)

Novel Computational Methods for Large Scale Genome Comparison

Doctorand: Todd James Treanger

Director: Xavier Messeguer

Resum: L'abundància de dades genòmiques disponibles actualment sobre una àmplia varietat d'espècies proporciona una oportunitat sense precedent per comparar i contrastar les històries evolutives dels organismes. La identificació de ADN homòlegs ha estat una metodologia fonamental en els estudis evolutius, i aquesta identificació s'ha fet utilitzant la tècnica bioinformàtica d'alineament de seqüències, tècnica que s'ha demostrat prou versàtil per a comparar organismes propers i distants. En el cas de dues seqüències l'alineament global i local òptim requereix temps i espai $O(m \times n)$ respecte de les longituds m, n de les seqüències, la qual cosa fa inviable la seva utilització per a seqüències genòmiques. Aquest coll d'ampolla empitjora si tenim k seqüències ja que el cost té un comportament exponencial respecte k , $O(nk)$, la qual cosa significa la inviabilitat de la tècnica també per a un nombre no gaire gran de seqüències curtes.

La finalitat d'aquesta tesi és el desenvolupament de nous algorismes i programari per a comparacions eficients globals i locals de múltiples genomes, i per a la aplicació d'aquest mètode per a l'estudi d'un cas de rellevància biològica. La investigació s'ha organitzat en base a d'aquest objectiu general en tres fases successives, específicament: (1) alineament de molts genomes d'espècies properes, (2) alineament local múltiple de repeticions entremesclades, i finalment, (3) l'aplicació d'aquestes tècniques a un estudi comparatiu dels genomes de la *Nesseria*.

En la primera fase, primer desenvolupem un algorisme eficient i una estructura de dades per facilitar la cerca de patrons comuns a varies seqüències, concretament per buscar las sub-seqüències úniques maximals (MUMs), obtenint millores significatives en temps i espai respecte els algorismes existents. Específicament, donats $S_1 \dots S_m$ genomes (on S_1 és la llargada del genoma més curt), podem trobar els MUMs entre totes les seqüències en temps lineal $O(|S_1| + \dots + |S_m|)$ i espai lineal $O(|S_1|)$. Aquest algorisme s'ha implementat dins de l'aplicació bioinformàtica interactiva, M-GCAT, que permet construir i visualitzar eficientment les semblances i diferències entre molts genomes d'espècies propers. En la segona fase, presentem un nou mètode computacional per a l'alineament múltiple local de repeticions entremesclades. Aquest mètode ofereix un nou enfocament per a l'extensió de molts patrons encadenats, ajuntant l'alineament múltiple global amb una prova de homologia basada en els Models Ocults de Markov (HMM).

Hem implementat el nostre mètode en l'aplicació procrastAlign. En la fase 3, utilitzant els programes obtinguts a les dues fases anteriors, realitzem un estudi dels genomes de l'espècie *Neisseria* per esbrinar la propagació d'uns elements repetitius al llarg de l'evolució de l'espècie i entendre per què els patògens importants del grup de *Neisseria* tenen un intercanvi sexual de DNA per transformació natural.

En conclusió, les contribucions globals d'aquesta tesi s'han centrat en el disseny de noves estructures de dades, algorismes i programari per a la investigació comparativa dels genomes que s'han demostrat molt eficients i s'han implementat en programari lliure, a més de la demostració de la seva utilitat per comparar i contrastar històries evolutives d'organismes relacionats.

Dia: Pendent de confirmació

Hora: Pendent de confirmació

Lloc: Pendent de confirmació

Contacte de premsa:
ilapiente@lsi.upc.edu