SOFT COMPUTING PER ANALISI FILOGENETICHE.

Giuseppe Pulcini, Francesco Riganti Fulginei, Alessandro Salvini

Dipartoimento di Elettronica Applicata, Università di Roma Tre
Via della Vasca Navale 84, 00146 Roma, Italy

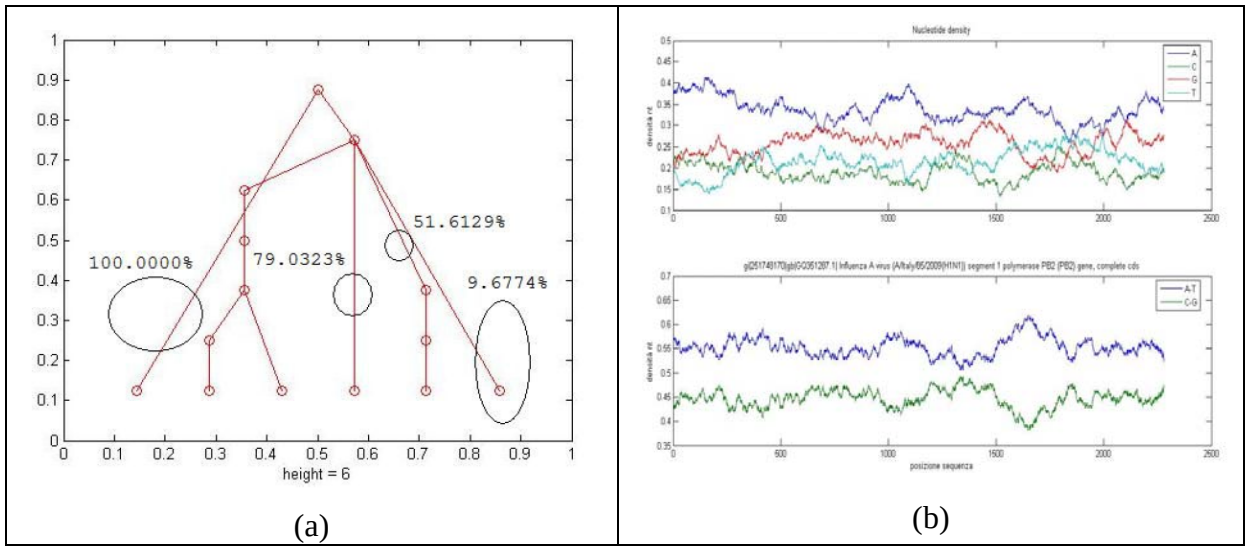
Il know-how sull'utilizzazione di tecniche di softcomputing che l'unità di Elettrotecnica di Roma Tre ha ascquisito negli ultimi anni occupandosi prevalentemente di problemi inversi e di ottimizzazione in campo elettromagnetico, ha consentito una serie di collaborazioni scientifiche nell'ambito dell'analisi delle sequenze biologiche. In particolare, si vuole qui riassumere l'attività di ricerca più recente in tale campo riguardante l'utilizzo di tre metaeuristiche per la ricostruzione di alberi filogenetici [1]: il Flok of Starling Optimization (FSO)[2], il Bacterial Chemotaxis Algorithm (BCA)[3] oltre al ben noto l'Algoritmo Genetico (AG). Il principio sul quale tali euristiche si fondano per risolvere il problema dell'albero minimo è il principio di "Massima Parsimonia" [4] che trova le sue radici nel concetto più generale di minima evoluzione. Le sequenze sulle quali abbiamo inferito con questi algoritmi sono quelle della SARS e una scelta tra le sequenze dell'influenza A/H1N1. Il metodo della massima parsimonia è probabilmente quello più diffuso e che viene maggiormente utilizzato tra quelli basati sull'analisi delle sequenze. Tale metodo consiste nell'associare ad ogni albero un certo punteggio, solitamente chiamato lunghezza, che rappresenta il numero minimo di sostituzioni necessarie per relazionare le sequenze in quel determinato modo, ovvero per definizione è l'albero con minori cambi di residui paralleli .L'albero (o gli alberi) più parsimonioso sarà quindi quello che avrà lunghezza minore. Quell'albero che alla fineper ogni sito informativo avrà lunghezza minore, considerando la somma di ogni singolo sito, sarà ilpiù parsimonioso. Formalizzando, in presenza di k siti, la lunghezza dell'albero L è data da $L = \sum_{i=1}^{n} l_i$. La descrizione dell'albero filogenetico da noi utilizzata è una matrice di incidenza nodi-lati, in cui è presente un 1 nella posizione i, j quando le due specie *i* e *j* sono legate da un ramo, contrario c'è il valore 0. Tale codifica risulta più conveniente rispetto a quelle in uso, come quella di Newick.

Come precedentemente detto, l'inferenza sulle sequenze è stata testata con l'utilizzo di 3 metaeuristiche FSO, BCA e AG. Il Flok of Starling Optimization è un algoritmo che prende ispirazione dai recenti studi naturalistici compiuti sul comportamento di un vero stormo di uccelli durante il volo. Il Bacterial Chemotaxis Algorithm è un algoritmo che si basa sul meccanismo biochimico che sta alla base del movimento dei batteri in un determinato ambiente durante la ricerca del cibo ed è in grado di ottenere ottimi risultati se sfruttato come algoritmo di ricerca locale [2]. L'algoritmo genetico è ispirato alla crescita di una popolazione di individui che attraverso la selezione naturale porta a migliorare le caratteristiche dei propri individui.

RISULTATI SPERIMENTALI

L'inferenza è stata fatta su sequenze genetiche della SARS, e su una selezione di sequenze dell'H1/N1. Per testare la validità delle singole euristiche è stato utilizzato un pool di 100 sequenze randomiche. La validazione è stata effettuata utilizzato il metodo bootstrap [5], trovando come valore di confidenza di ogni nodo in alcuni casi anche del 100% (Fig.1),

gli alberi ricampionati sono stati 62. Alberi 90-parsimonia presentano la metà dei collegamenti monofiletici all'interno dell'albero con valori di confidenza maggiori del 50%. Il costrutto risponde in tempi moderatamente piccoli in media di 8'. Nelle simulazioni con 100 sequenze random, il tempo impiegato è stato di circa 12 ore. In generale le euristiche senza collective behavior performano meglio. La spiegazione risiede nel fatto che lo spazio delle soluzioni ha delle zone si singolarità molto ampie che rendono difficoltosa la dinamica collettiva.



(a) (b)

Figura 1: (a) Albero SARS con livelli di frequenza segnati per alcuni lati. (b) Frequenza nucleotidica di una sequenza H1/N1.

Referenze

[1] LJ Billera, SP Holmes, K Vogtmann. Geometry of the Space of Phylogenetic Trees. Advances in Applied Mathematics, Vol. 29, Issue 1, July 2002, Page 136.

[2] F. R. Fulginei, A. Salvini, Hysteresis model identification by the Flock-of-Starlings Optimization, Int. Journal of applied Electromagnetics and Mechanics, IOS Press, vol. 30, No. 3/4 , 2009, pp.321-331.

[3] S.D. Muller, J. Marchetto, S. Airaghi, P. Kournoutsakos, Optimization based on Bacterial Chemotaxis", IEEE Trans. on Evolutionary Computation, Vol. 6, No. 1, February 2002, pp. 16-29.

[4] David A. Bader, Vaddadi P. Chandu,Mi Yan. ExactMP: An Efficient Parallel Exact Solver for Phylogenetic Tree Reconstruction Using Maximum Parsimony. International Conference on Parallel Processing,IEEE. 2006.

[5] Hillis Dm, Bull JJ. An Empirical-test of bootstrapping as a method for assensing confidence in phylogenetic analysis. Systematic Biology 1993, Vol 42,2.