Tesi di Laurea Specialistica

Elaborazione di dati bioinformatici attraverso l’uso di Particle Swarm Optimization

Candidato
Davide Chicco

Relatori
Francesco Masulli, Giuseppe Russo

Correlatore
Giorgio Delzanno
Indice

1 Sommario 5

2 Introduzione 6

3 Elementi di biologia molecolare 8
   3.1 Acidi nucleici ................................. 8
      3.1.1 Basi azotate .......................... 10
      3.1.2 Trascrizione ......................... 11
   3.2 microRNA .................................. 11
   3.3 Allineamento di sequenze .................. 13
      3.3.1 Interpretazione dell’allineamento .... 13
      3.3.2 Metodi d’allineamento ............... 14
   3.4 Energia libera ............................. 15
      3.4.1 Definizione ........................... 16
      3.4.2 Minimo energetico .................... 17
      3.4.3 Energia libera e folding ............. 17
      3.4.4 Esempio di modellazione della struttura proteica 18

4 Introduzione alla predizione di geni obiettivo dei microRNA 20
   4.1 Software disponibili ....................... 20
      4.1.1 Sensibilità e specificità .......... 24
   4.2 miRanda .................................. 25
   4.3 Valutazione dell’omologia ................ 25
      4.3.1 L’algoritmo di Smith-Waterman ...... 26
         Come funziona .......................... 26
         Un esempio ............................. 27
         Idea generale .......................... 28
   4.4 Minimizzazione dell’energia libera .... 29
   4.5 Conservazione evolutiva .................. 29
   4.6 Euristie di miRanda ....................... 29
   4.7 Mirta .................................... 30
      4.7.1 Miglioramenti rispetto a miRanda .... 30
         Simple Genetic Algorithm (SGA) ...... 31
Funzione d'ottimizzazione .......................... 32
4.7.2 Insieme dei dati considerati ................... 34

5 Particle Swarm Optimization (PSO) 35
  5.1 Algoritmi evolutivi ............................... 35
    5.1.1 Implementazione dei processi biologici .......... 36
    5.1.2 Tecniche di calcolo evolutivo .................. 37
  5.2 PSO ............................................. 37
    5.2.1 Idea generale ................................ 38
    5.2.2 L'algoritmo .................................. 39
    5.2.3 Esempio ...................................... 40
     Funzione 1 di De Jong ................................ 40
     Funzione 2 di De Jong ................................ 41
     Funzione 3 di De Jong ................................ 42
     Funzione 4 di De Jong ................................ 44
     Funzione 5 di De Jong ................................ 45
     Funzione di Rosenbrock standard ...................... 46
     Funzione di Rosenbrock modificata ..................... 47

6 Metodo proposto 49
  6.1 Parametri ....................................... 50
  6.2 Funzione d'ottimizzazione ....................... 51
  6.3 Insieme dei dati considerati .................... 52

7 Implementazione 54
  7.1 GLibrary ....................................... 54
    Tipi primitivi ................................... 55
    Macro ............................................ 55
    Memoria ......................................... 56
    Nel programma .................................... 56
  7.2 Vienna Library ................................... 56
  7.3 Strutture dati ................................... 57
  7.4 Funzione d'ottimizzazione ....................... 58
  7.5 Main ........................................... 61

8 Risultati ottenuti 64

9 Conclusioni 70
  9.1 Sviluppi futuri ................................. 71
    9.1.1 Alternative al PSO ............................ 71
    9.1.2 Versioni accelerate dell'algoritmo di Smith-Waterman:  
      GPU ............................................. 72
    9.1.3 Possibilità di parallelizzazione ................. 72

10 Appendice A: Manuale d’uso del programma 74
Ringraziamenti

Questo progetto è nato all’interno d’una collaborazione tra il gruppo di Machine Learning e Soft Computing del DISI ed il prof. Giuseppe Russo dello Sbarro Institute for Cancer Research and Molecular Medicine, presso il Center for Biotechnology alla Temple University in Filadelfia (Pennsylvania, Stati Uniti).

Un particolare ringraziamento va, oltre ai prof. Giuseppe Russo e Francesco Masulli, al prof. Stefano Rovetta ed alla dott.ssa Maura Monville del DISI per l’aiuto datomi.

Un ringraziamento anche al collega Alessandro Parini per avermi messo a disposizione la versione precedente del software sui cui ho lavorato.

Chiunque volesse contattarmi, anche in futuro, per avere informazioni sul presente progetto, è invitato a scrivermi all’indirizzo mail davide.chicco@gmail.com
I microRNA (o miRNA) sono molecole di RNA non codificante che regolano negativamente l’espressione di geni coinvolti in svariati pathways metabolici che possono promuovere lo sviluppo del cancro, del morbo di Parkinson, oltre ad infezioni virali. Esiste un’area della bioinformatica che si pone come obiettivo quello dell’identificazione, attraverso metodi computazionali, di geni obiettivo (target) di queste molecole di microRNA.

I risultati dei programmi di predizione disponibili al momento, come ad esempio il software miRanda sviluppato da A.J. Enright [1] e Mirta [2] [3] [4], programma basato su miRanda che ne migliora notevolmente alcuni aspetti, devono essere validati biologicamente per essere considerati affidabili.


In questo modo è stato possibile migliorare la consistenza tra geni obiettivo (targets) predetti e targets validati biologicamente rispetto alla versione originale di miRanda [1].
Capitolo 2

Introduzione

Un problema importante all’interno della bioinformatica è quello della predizioni di siti obiettivo (target) all’interno di sequenze di RNA (acido ribonucleico).

I microRNA sono una classe di RNA non codificanti (ncRNA) di circa 18-25 nucleotidi di lunghezza, in grado di partecipare a processi di accoppiamento di basi (base-pairing, anche con complementarità imperfetta negli animali) con trascritti (transcripts) di geni (detti anche geni bersaglio-targets) che codificano proteine, generalmente all’interno della regione 3’-UTR (“un-translated region”, una sezione finale non tradotta dell’RNA messaggero).


Allo stato attuale, gli strumenti di predizione disponibili offrono risultati discordanti, in particolare alcuni geni target validati biologicamente non sono predetti da nessun metodo. La validazione in laboratorio di targets predetti non è ancora esaustiva perché le procedure biologiche non possono trovare i siti dei targets dei microRNA.

Per questo obiettivo, negli ultimi anni sono stati implementati e resi disponibili diversi software bioinformatici dedicati alla ricerca di siti targets nei microRNA. Uno dei più diffusi al momento è miRanda [1], sviluppato nel 2003 da A. J. Enright ed altri, del laboratorio di biologia computazionale del Memorial Sloan-Kettering Center di New York City, negli Stati Uniti.
Questo programma s’è dimostrato utile ed affidabile nel corso degli anni, anche se non definitivo.
In particolare, è risultato evidente che alcuni miglioramenti potessero essere raggiunti attraverso una scelta più accurata dei parametri da usare, e in particolare attraverso l’uso di algoritmi genetici [2], o altre tecniche d’ottimizzazione.
L’obiettivo della mia tesi è la predizione computazionale di targets dei microRNA, in particolare quelli animali che formano appaiamenti di basi con complementarità imperfetta, attraverso una modifica al programma miRanda e la scelta dei parametri attraverso l’algoritmo d’ottimizzazione PSO (Particle Swarm Optimization [9]).
Nella prima parte della tesi verrà prima fatto un accenno alla problematica biologica, e poi verranno spiegati i metodi proposti e la loro implementazione software. Infine, verranno analizzati i risultati ottenuti e presentate alcune riflessioni conclusive sul lavoro svolto.
In dettaglio, i capitoli 3 e 4 forniranno un’introduzione al problema su cui ho lavorato: nel capitolo 3 vedremo un’introduzione a qualche concetto biologico di base. Nel capitolo 4 vedremo a grandi linee come funziona il programma per il rilevamento di siti targets per le sequenze genomiche. Nel capitolo 5 daremo uno sguardo d’insieme al PSO, l’algoritmo evolutivo usato qui per la rifinitura dei parametri. Nel capitolo 6 vedremo come è stato impostato il metodo proposto. I dettagli implementativi sono riportati nel capitolo successivo (capitolo 7). Seguiranno un capitolo sui risultati ottenuti (capitolo 8) ed uno sulle conclusioni sul lavoro svolto e gli sviluppi futuri (capitolo 9).

Infine, un’appendice dedicata al manuale d’uso del programma costituisce il capitolo 10.