Búsqueda de Motivos Biológicos en Secuencias de ADN de origen Ambiental

Luis Garreta¹, Patricia Vélez², Pedro Moreno³, e Irene Tischer⁴

Centro de Excelencia en Genómica Y Bioinformática De Ambientes Extremos -GeBiX Grupo de Bioinformática y Biocomputación Universidad del Valle -Cali-Colombia







Agenda

Introducción

2 El Enfoque

3 Resultados

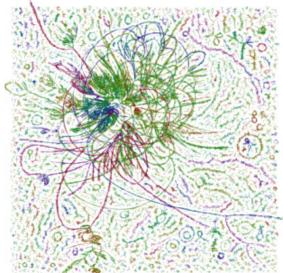
Introducción



GEBIX Centro Colombiano de Genómica y Bioinformática de Ambientes Extremos

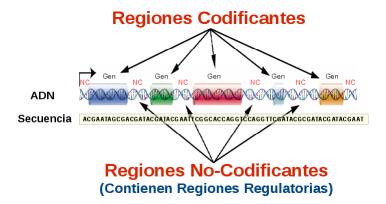


Metagenómas

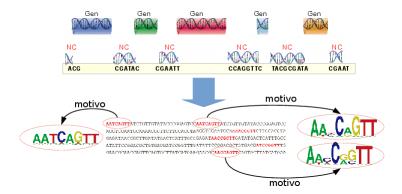




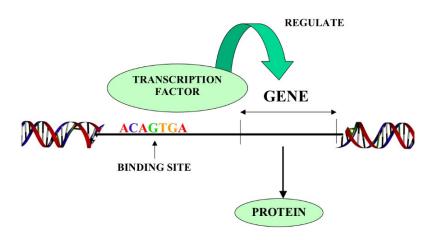
Regiones No-Codificantes



Extracción de las Regiones No-Codificantes



Sitios de Enlace de Factores de Transcripción



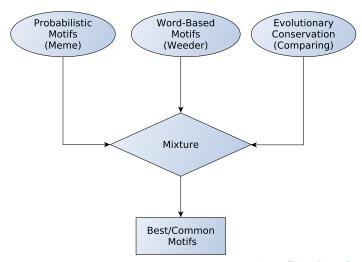
Búsqueda de Motivos en Secuencias de ADN

qtqaqaqatqaqcqaqacaqaaaqaqaqaqaqtqcaccaqcqaqccqqqqcaqqaaqaqqtttcqccaccqqaqcqqcccqqcqacqcttacaqcttcccc tqcccttcccqtcqqtcqqccqccaqccqccqcaqccctcqqcctqcacqccaccqqccccqctcccqqaqcccaqcqccqcqqcqaqccqcaqccqccqccq gtaaggeggegeegeegeeggeeaceggegeeetgegetteeeteegeeggetgeggeeatggegegetgaetggeetgaetggeeeggeeeggeegegeteeeg ggaggccgaacgcggactcgccacccggcttcagaatggcagaagatgatccatatttgggaaggcctgaacaaatgtttcatttggatccttctttgactcatacaa tatttaatccaqaaqtatttcaaccacaqatqqcactqccaacaqatqqcccataccttcaaatattaqaqcaacctaaacaqaqaqqatttcqtttccqttatqtat gtgaaggcccatcccatggtggactacctggtgcctctagtgaaaagaacaagaagtcttaccctcaggtcaaaatctgcaactatgtgggaccagcaaaggttattg ttcagttggtcacaaatggaaaaaatatccacctgcatgcccacagcctggtgggaaaacactgtgaggatgggatctgcactgtaactgctggacccaaggacatgg tgqtcqqcttcqcaaacctqqqtatacttcatqtqacaaaqaaaaaaqtatttqaaacactqqaaqcacqaatqacaqaqqcqtqtataaqqqqctataatcctqqac tettqqtqcaccetqacettqcctatttqcaaqcaqaaqqtqqaqqqqaccqqcaqetqqqaqatcqqqaaaaaqqqctaatccqccaaqcaqctetqcaqcaqacca aggagatqqacctcagcqtqqtqcqqctcatqtttacagcttttcttccqqatagcactqqcagcttcacaaqqcqcctqqaacccqtqqtatcaqacqccatctatq aagatgacatccagattcgattttatgaagaggaagaaaatggtggagtctgggaaggatttggagatttttcccccacagatgttcatagacaatttgccattgtct tcaaaactccaaaqtataaaqatattaatattacaaaaccaqcctctqtqtttqtccaqcttcqqaqqaaatctqacttqqaaactaqtqaaccaaaacctttcctct actatectqaaateaaaqataaaqaaqtqcaqaqqaaacqtcaqaaqetcatqeecaattttteqqataqttteqqeqqtqqtaqtqqtqqtqqtqqaqetqqaqqeq qaqqcatqtttqqtaqtqqcqqtqqaqqqqqqcactqqaaqtacaqqtccaqqqttataqcttcccacactatqqatttcctacttatqqtqqattactttccatc ctggaactactaatctaatgctgggatgaagcatggaaccatggacactgaatctaaaaaggaccctgaaggttgtgacaaaagtgatgacaaaaacactgtaaacc tetttqqqaaaqttattqaaaccacaqaqcaaqatcaqqaqeccaqcqaqqccaccqttqqqaatqqtqaqqtcactetaacqtatqcaacaqqaacaaaaqaaqa gtgctggagttcaggataacctctttctagagaaggctatgcagcttgcaaagaggcatgccaatgccattttcaactacgcggtgacaggagacgtgaagatgctgc tggccgtccagcgccatctcactgctgtgcaggatgagaatggggacagtgtcttacacttagcatcatccaccttcatctcaacttgtgagggatctactagaag tcacatctggtttgatttctgatgacattatcaacatgagaaatgatctgtaccagaggccttgcacttggcagtgatcactaggcaggaagatgtggtggaggatt tgctgagggctggggccgacctgagccttctggaccgcttgggtaactctgtttk.cacctagctgccaaagaaggacatgataaagttctcagtatcttactcaagc acaaaaaqqcaqcactacttcttqaccaccccaacqqqqacqqtctqaatqccatqctatqccatqatqaqcaataqcctqccq;qtttqctqctqqtqqccq agctgggagagggtccaccaggctggcagatcttctcaaagcagcaggagcagatc ccctd Itgcaggagaggatgaaggagttgtgcctggaaccacgcctctagatatggccacca actac tgcag 4 ctggcgcagaaattaggtctggggatacttaataatgccttccggctgagtcctg ctcct a agetggtggaggeetgagaeaatgggetaeacegaageaattgaagtgatee LGCC coccepectecacaaggcagcaaatagacgagctccgagacagtgacagtgtct qcqac onccagtggtgcctcactgctaactctcaacaaaatgccccatgattatgggcagg aagga (ccaaa gaccttttcaacttggcttcctttcttggttcataaatgaattttagtttggttca adtccaccaggctggcagctctggatgaggttgcttactaagctttgccagctgctgctggatcacagctgctttctgt cttad tqtca :qqtcqccacctqqcattccttctqaccacaqcatcattttqcattcaaattaaqqq aaaaaaaaaatacttgtcaatatttaaacatggttacaatcattgctgaaaatggtattttcccccttttctgcattttgctattgtaaatatgttttttagatcaaa

Enfoque

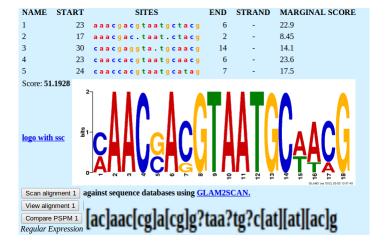
Múltiples Fuentes:

Mezcla de Expertos





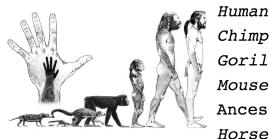
Word-based Method



Probabilistic Method

```
a HEM13
         CCCATTGTTCTC
         TTTCTGGTTCTC
  HEM13
  HEM13
         TCAATTGTTTAG
  ANB1
         CTCATTGTTGTC
  ANR1
         TCCATTGTTCTC
                                        Scan direction
  ANB1
         CCTATTGTTCTC
                         tgcggaatgcgctttgtttttatcaaaaaaaacacccgcacatgcatcagtgtcatat
  ANB1
         TCCATTGTTCGT
  ROX1
         CCAATTGTTTTG
                                                                         MATCH
                                            6
b
         YCHATTGTTCTC
                                -8
                                                                             Yes
                                                       ≥ threshold
C
        002700000010
                                                                             No
         464100000505
         000001800112
                                                                       NO MATCH
         422087088261
                             2+4+2-3+3-3=5
```

Evolutionary Approach



Human ATGAACGCATGC
Chimp. ATGCACGCATGC
Gorilla ATGCATGCATGC
Mouse ATGCATGCATGC
Ancestor ATGCATGCACGC

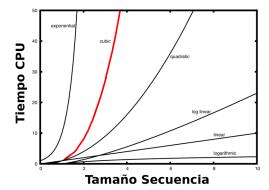
ATGCATGCACGC

Resultados

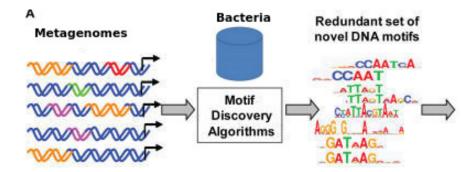


Algoritmos no Funcionan Bien con Secuencias Largas

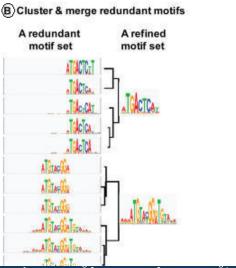
- Word-based Method: Weeder
- Probabilistic Method: Meme



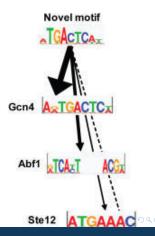
Enfoque Evolutivo - Parte A



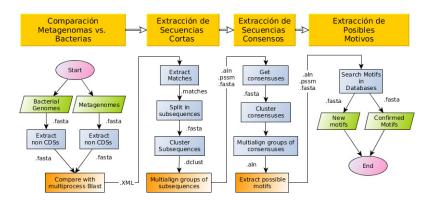
Enfoque Evolutivo - Parte B



C Relating novel motifs to known TF binding sites



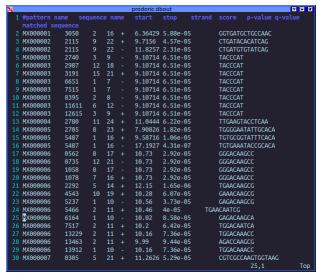
Pipeline Para Búsqueda de Motivos en Regiones No-Codificantes



Bases de Datos para Validar los Motivos

- Prokaryotes (Prodoric Release 8.9): Prokaryotic motifs from PRODORIC downloaded from http://prodoric.tu-bs.de/msearch.php.
- Prokaryotes (RegTransBase v4): Manually curated prokaryotic motifs from RegTransBase; downloaded from http://regtransbase.lbl.gov/cgibin/regtransbase?page=downloads.
- E. coli **(DPINTERACT): Footprinting databases based on DPINTERACT; downloaded from http://arep.med.harvard.edu/ecoli_matrices.

Resultados Preliminares



Gracias!