BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 14 2010, pages 1777–1778
doi:10.1093/bioinformatics/btq258

Genome analysis

Advance Access publication May 25, 2010

COMPASSS (COMplex PAttern of Sequence Search Software), a simple and effective tool for mining complex motifs in whole genomes

Giuseppe Maccari[1,*], Federica Gemignani[2,†] and Stefano Landi[2,†,*]

[1]Laboratory of Microbiology and Genetics – Ospedale di Circolo e Fondazione Macchi, University of Insubria, Viale Borri 57 21100 Varese, Italy and [2]Department of Biology – Genetics, University of Pisa, via Derna 1, 56126 Pisa, Italy

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The complete sequencing of the human genome shows that only 1% of the entire genome encodes for proteins. The major part of the genome is made up of non-coding DNA, regulatory elements and *junk DNA*. Transcriptional regulation plays a central role in a multitude of critical cellular processes and responses, and it is a central force in the development and differentiation of multicellular organisms. Identifying regulatory elements is one of the major tasks in this challenge. To accomplish this task, we developed a solid and simple suite that allows direct access to genomic database and immediate result check. We introduce COMPASSS (COMplex PAttern of Sequence Search Software), a simple and effective tool for motif search in entire genomes. Motifs can be partially degenerated and interrupted by spacers of variable length.

Results: We demonstrate through real biological data mining the simplicity and robustness of this tool. The test was performed on two well-known protein domains and a highly variable *cis*-acting element. COMPASSS successfully identifies both protein domains and *cis*-acting semi-conserved elements.

Availability: The COMPASSS suite is available for Windows free of charge from our web sites: compasss.sourceforge.net/; www.stefanolandi.eu/

Contact: gpmaccari@gmail.com; slandi@biologia.unipi.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 18, 2009; revised on May 12, 2010; accepted on May 14, 2010

1 INTRODUCTION

Understanding the significance of unannotated DNA sequences from genome projects is one of the major challenges, nowadays. Interpretation of the genetic code allowed to discover over 30 000 protein-coding genes in the human genome (International Human Genome Sequencing Consortium, 2004). However, the major part of the genomes is not made up of genes encoding for proteins (Venter *et al*., 2001). The genetic code can be transcribed in non-coding RNA or it can be constituted of *cis*-acting regulatory sequences such as promoters, enhancers and other elements that affect gene expression and alternative splicing (Balakirev and Ayala, 2003; Tress *et al*., 2007). Moreover, most of the genomes are constituted by intergenic sequences that may have a regulatory role, but are defined as 'junk DNA' because we are unable to understand its function, yet.

Nowadays, the availability of a wide range of genome sequences has greatly accelerated research to decipher the genomic regulatory code. To clarify the cryptic nature of the human genome, the ENCODE (the ENCyclopedia Of DNA Elements) research consortium has been launched. Functional elements in the human genome will be shared in a public database (The ENCODE Project Consortium, 2007). The pilot phase of the project tested and compared existing methods to analyze a small portion of the human genome sequence (1%). Now the project has been scaled to the entire human genome. *In silico* motif search within DNA sequences is one of the most challenging problems in molecular biology and computer science and there is a large body of work done in developing search algorithms. However, each of these approaches has its own strengths and weaknesses, while the quality of the detected motifs may vary from case to case (Das and Dai, 2007).

In this work, we propose a simple and effective tool named COMPASSS (COMplex Pattern of Sequence Search Software) that allows mining whole genomes for the presence of complex elements. The tool is user-friendly, flexible, not resource intensive and capable of browse whole genomes in minutes. COMPASSS applies the well-known Wu-Manber multiple pattern matching algorithm to exhaustively search for motifs in the entire sequence. Input elements can be simple conserved sequences (as in BLAST engine), partially or highly degenerated sequences, or even multiple degenerated sequences, joined by poorly conserved variable length spacers. We successfully tested COMPASSS against three experimentally validated complex patterns, providing evidences that the tool successfully identifies both protein domains as well as *cis*-acting semi-conserved elements.

2 METHODS

2.1 Motif definition

Input motifs can be composed of one or more *patterns* interspaced by zero or more *spacers*. A pattern can be specified by regular bases (A, T, G, C, U) as well as IUPAC ambiguous nucleotides or *positional weight matrix* (PWM). A spacer is defined by a fixed length l and a range r; the final spacer length can vary from $l-r$ and $l+r$.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Authors.

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

2.2 Primary pattern definition

Before the searching phase, a primary pattern is defined. COMPASSS starts the motif search from the most complex pattern specified within the input motif. Pattern complexity is proportional to its length and ambiguity. A longer pattern is less likely to match the input sequence:

$$1-\prod_{x=1}^{K}\frac{1}{R_x}$$

where K is the pattern length and $R_x$is the number of possible matching symbols at pattern position x.

2.3 Motif search

Our approach uses a modified implementation of the *Wu-Manber* algorithm (Wu and Manber, 1994) to search for the desired motif within the whole input sequence. This method has been chosen for the ability to search in the input sequence for multiple patterns at the same time and for no preprocessing phase on the input string. The last point has great importance; the aim was to create a flexible and robust program for the search of motif into whole genomes. In the preprocessing phase, the original *Wu-Manber* algorithm builds three tables to speed up the text processing. The SHIFT table is used to determinate how many characters in the text can be shifted when the text is scanned. The HASH and PREFIX tables are used when a match is found and to verify the match. The algorithm reads blocks of characters and search for a match in the shift table. If a correspondence is found, the algorithm checks the entire string for a match. Otherwise the text is shifted. In this implementation of the algorithm a fourth table is added, the FOUND buffer. The found buffer is a circular buffer of the same size of the defined motif that shifts with the input sequence to keep track of the latest patterns found. When a match is found, the FOUND buffer is checked for the previous pattern at the defined position. If the pattern is found, the FOUND buffer is updated. If the last item of the motif is found, a new result is added. This approach greatly reduces memory usage and allocation of incomplete results.

For each found motif, a score is given. The total score of a found motif is given by the sum of every pattern score (S_i) normalized with its spacer score:

$$\sum_{i=1}^{n} S_i\left(1-\frac{|ld_i-le_i|}{r_i}\right) \quad \text{where, for PWM:} \quad S_i=\prod_{j=1}^{m} p_j$$

The pattern score S_i is 1 for IUPAC input sequences and the product of each positional score (P_j) for PWM input patterns. The spacer score is defined by the difference between the defined spacer length (ld) and the actual found value (le) in absolute value, normalized by the given range (r).

2.4 Database mining

The COMPASSS suite has built-in database mining capability. The motif search process can be accomplished on local files or on NCBI nucleotide database entries through NCBI entrez utility. Moreover, COMPASSS can search for motif on specific genomic regions, without the need to download entire sequences. This task is accomplished thanks to the DAS protocol (www.biodas.org) and the MySQL access to ENSEMBL servers (http://www.ensembl.org/info/data/mysql.html). Test data were downloaded via ENTREZ utilities at NCBI database. For the DAS search, the ensemble server has been used (http://www.ensembl.org/das/sources).

3 RESULTS

To evaluate the effectiveness of the algorithm, tests were performed on actual biological data. In particular, we focused our attention on two semi-conserved protein domains and a *cis*-acting element: the human globin gene family, the mammalian POU-domain transcription factors, and the *TP53* responsive elements. In order to define the signature of each input motif, CLUSTALW algorithm was run and a multiple alignment of known target regions was obtained. Thus, conserved and degenerated bases were identified and input motif was defined.

The human globin gene family input motif was composed of three degenerated sequences, separated by two conserved spacers (CTGCACKS<13,2>GTGGAYCC<12,0>GGTG) and designed to recognize the second exon of the alpha and beta locus. The software successfully identifies both the alpha and beta globin genes.

The POU domain motif consists of a bipartite motif on the C-terminal region of the POUH domain (Ryan and Rosenfeld, 1997). The two degenerated sequences of the POU motif are separated by a spacer of a fixed length of 5 bp. The input motif was MGNGTBTGGTTYTGYAA<5,1>GVCARR. A search against the entire *Homo sapiens* genome results in a complete list of all known members of the POU domain family, plus two genes with high affinity and homology with the POU domain family.

The tetrameric p53 protein binds to two repeats of a consensus DNA sequence composed by two conserved pairs each arranged as inverted complement elements separated by a variable spacer region (El-Deiry *et al*., 1992; Levine, 1997). With the COMPASSS suite, we searched for all p53 regulatory motives throughout the entire human genome using the input data RRRCWWGYYY<7,7>NRRCWWGYYN. The software identified most of the experimentally validated p53 responsive elements. Detailed results can be found in Supplementary Material.

4 CONCLUSIONS

In this article, we showed that COMPASSS is an efficient and powerful tool for capturing most of the real motifs with proven biological activity. In contrast with other motif-search tools, COMPASSS is efficient, fast and not resource intensive. The suite can be used for mining *cis*-acting elements, including promoters, splice sites or regulatory regions.

Funding: Associazione Italiana Ricerca Cancro (AIRC), Investigator Grant 2008.

Conflict of Interest: none declared.

REFERENCES

Balakirev,E.S. and Ayala,F.J. (2003) Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.*, **37**,123–151.

Das,M.K. and Dai,H.K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8** (Suppl. 7), S21.

El-Deiry,W.S. *et al.* (1992) Definition of a consensus binding site for p53. *Nat. Genet.*, **1**, 45–49.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Levine,A.J. (1997) p53, the cellular gatekeeper for growth and division *Cell*, **88**, 323–331.

Ryan,A.K. and Rosenfeld,M.G. (1997) POU domain family values: flexibility, partnerships, and developmental codes. *Genes Dev.*, **11**, 1207–1225.

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Tress,M.L. *et al*. (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.

Venter,J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wu,S. and Manber,U. (1994) A fast algorithm for multi-pattern searching. *Technical Report TR94-17*, Department of Computer Science, University of Arizona.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016