ABSTRACT

Motivation: Many analyses in modern biological research are based on comparisons between biological sequences, resulting in functional, evolutionary and structural inferences. When large numbers of sequences are compared, heuristics are often used resulting in a certain lack of accuracy. In order to improve and validate results of such comparisons, we have performed radical all-against-all comparisons of 4 million protein sequences belonging to the RefSeq database, using an implementation of the Smith–Waterman algorithm. This extremely intensive computational approach was made possible with the help of World Community Grid™, through the Genome Comparison Project. The resulting database, ProteinWorldDB, which contains coordinates of pairwise protein alignments and their respective scores, is now made available. Users can download, compare and analyze the results, filtered by genomes, protein functions or clusters. ProteinWorldDB is integrated with annotations derived from Swiss-Prot, Pfam, KEGG, NCBI Taxonomy database and gene ontology. The database is a unique and valuable asset, representing a major effort to create a reliable and consistent dataset of cross-comparisons of the whole protein content encoded in hundreds of completely sequenced genomes using a rigorous dynamic programming approach.

Availability: The database can be accessed through http://proteinworlddb.org

Contact: otto@ficruz.br

Received on April 25, 2009; revised on January 6, 2010; accepted on January 9, 2010

1 INTRODUCTION

The assignment of biological function predictions and structural features to raw sequence data is typically accomplished by comparing them either to predicted protein sequences or to the corresponding genes. This information is stored in several primary public databases, such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) or EMBL-Bank (http://www.ebi.ac.uk/embl/). However, annotations are often incomplete, based on non-standardized nomenclature or might have no value when inferred from previous incorrectly annotated sequences. Hence, secondary databases such as Swiss-Prot (http://www.expasy.ch/sprot/), PFAM (http://pfam.sanger.ac.uk) or KEGG (http://www.genome.ad.jp/kegg), to mention only a few, have been implemented to analyze specific functional aspects and to improve the annotation procedures and results.

Dynamic programming algorithms, or a fast approximation, have been successfully applied to biological sequence comparison for decades, and this class of algorithms comprises the heart of many well-known sequence alignment programs (Batzoglou, 2005). However, because of their quadratic time complexity, rigorous dynamic programming algorithms are usually not suitable for the comparison of a large set of sequences against a database, as they demand exceptionally huge computational power and are very time consuming. For this reason, sequence comparisons are generally performed by heuristics like BLAST (Altschul et al., 1997) and FASTA (Pearson, 1990), which have proved to be quite effective and significantly faster than the dynamic programming algorithms. However, in many instances, these comparisons might lack accuracy, as these heuristics do not guarantee to find a mathematically optimal alignment (Pearson, 1990), therefore affecting all subsequent analytical steps. The Genome Comparison Project (GCP) (http://www.dbbm.fiocruz.br/GenomeComparison) aims to compare protein information on a genomic scale to improve the quality and interpretation of biological data and our understanding of biological systems and their interactions. Stringent comparisons were obtained after the application of the Smith–Waterman (SW) algorithm (Smith and Waterman, 1981) in a pairwise manner to all predicted proteins encoded in both completely sequenced and unfinished genomes available in the public database RefSeq (version 21). The project represents a joint effort involving Fiocruz, PUC-Rio and IBM®, and was executed through World Community Grid™ (WCG), a computational grid on a global scale. We present here the outcome of this joint effort, the ProteinWorldDB, which represents a major effort to create a reliable and consistent dataset of cross-comparisons of the whole protein content encoded in hundreds of completely sequenced genomes using a rigorous dynamic programming approach.
The core of the ProteinWorldDB comprises the results of all pairwise functional classification (Swiss-Prot/TrEMBL), domain and protein family taxonomic information (NCBI Taxonomy database), gene ontology (GO), several third-party annotations, including gene and protein features (RefSeq), were generated and stored.

and similar regions, percentage of identity, number of gaps, raw and bit scores containing sequence identifiers, alignment length, coordinates of the most

(iii) maximum likelihood estimates of Lambda and K and the (iv) Altschul–Sims model implemented in the SSEARCH algorithm: (i) a weighted regression pair, taking the entire dataset into account, using four different mathematical models implemented in the SSEARCH algorithm. (i) a weighted regression method and then compared, after satisfying a certain threshold (Rattie et al., 2008). As previous studies have shown (Pearson, 1990; Uchiyama, 2007), the latter strategy is not guaranteed to find all hits. One should keep in mind that false positives are expected to be found with an E-value threshold of \( E \leq 0.001 \), as millions of comparisons were done. Nevertheless, function transfer and homology inference should not rely on E-value thresholds alone, since the fraction of identical positions between a pair of sequences, as well as the extension of their overlapping area, among several other sequence properties, play an important role in functional and evolutionary predictions based on sequence similarity (Boekhorst and Snel, 2007; Rost, 2002; Tian and Skolnick, 2003).

Valuable and unique information can be retrieved from ProteinWorldDB. For instance, queries could include: (i) individual or groups of proteins and their similarities with other entries based on the SW algorithm; (ii) download of subsets of the comparison data, i.e. related proteins shared by two particular species (inferred orthologs) or related proteins present in the same organism (inferred paralogs); (iii) genes that are exclusive of a particular species, i.e. taxonomically restricted (unique) genes; (iv) groups of related proteins for particular species using a protein of interest or a shared biological function as reference; and (v) comparison of different annotations for each entry. The ProteinWorldDB will, no doubt, contribute to improve annotation, to studies on genome and protein family evolution and in many other research aspects.

### 3.1 Further work
At this moment, the database contains similarity information using an E-value cutoff of \( 10^{-3} \). Later on, we will add additional results up to an E-value of one, and comparisons of an experimental set of open reading frames, which have not been predicted as coding. Datasets comprising different phylogenetic experiments, phylogenomics and horizontal gene transfer are in construction. Also, an update can be envisaged with the WCG to compute all the genomes that were included in RefSeq since the end of our experiments. In the future, we hope to develop automatic algorithms to scan differences in annotation between third-party databases, evaluate the confidence of the annotations, add a wiki-like annotation support system, allowing other groups to include their expertise in the database, as well as refine the interface in order to allow more complex queries.

### 3 ACKNOWLEDGEMENTS
We wish to thank IBM®®, World Community Grid™, Rede Fiocruz, Plataforma de Bionformatica PDTIS, Coord. Diao de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ),
Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their support.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. Methods Enzymol., 266, 460–480.

Altschul,S.F. et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

Batzoglou,S. (2005) The many faces of sequence alignment. Brief. Bioinform., 6, 6–22.

Brockhart,J. and Snel,B. (2007) Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. BMC Bioinformatics, 8, 356.

Kanehisa,M. et al. (2006) From genomes to chemical genomics: new developments in kegg. Nucleic Acids Res., 34, 354–357.

Otto,T.D. et al. (2008) AnEnPi: identification and annotation of analogous enzymes. BMC Bioinformatics, 9, 544.

Pearson,W. (1990) Rapid and sensitive sequence comparison with fastp and fasta. Methods Enzymol., 183, 63–88.

Pearson,W. (1998) Empirical statistical estimates for sequence similarity searches. J. Mol. Biol., 276, 71–84.

Pearson,W. and Sierk,M.L. (2005) The limits of protein sequence comparison? Curr. Opin. Struct. Biol., 15, 254–260.

Rattei,T. et al. (2008) SIMAP structuring the network of protein similarities. Nucleic Acids Res., 36, D289–D292.

Rost,B. (2002) Enzyme function less conserved than anticipated. J. Mol. Biol., 318, 595–606.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosquences. Adv. Appl. Math., 2, 482–489.

Tom,W and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol., 333, 863–882.

Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res., 35, D343–D346.