BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 5 2010, pages 705–707
doi:10.1093/bioinformatics/btq011

Databases and ontologies

Advance Access publication January 19, 2010

ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes

Thomas Dan Otto[1,2,*], Marcos Catanho[1], Cristian Tristão[3], Márcia Bezerra[3], Renan Mathias Fernandes[4], Guilherme Steinberger Elias[4], Alexandre Capeletto Scaglia[4], Bill Bovermann[5], Viktors Berstis[5], Sergio Lifschitz[3], Antonio Basílio de Miranda[1] and Wim Degrave[1]

[1]Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil, [2]Pathogen Genomics, Wellcome Trust Genome Campus, Hinxton, UK, [3]Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, [4]IBM Brasil, Hortolândia, São Paulo, Brazil and [5]IBM, Austin, TX, USA

Associate Editor: Alfonso Valencia

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@fiocruz.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.

*To whom correspondence should be addressed.

2 METHODS

The core of the ProteinWorldDB comprises the results of all pairwise comparisons accomplished by the GCP. Briefly, a set of 3 812 663 proteins from RefSeq version 21—consisting of all predicted proteins encoded in 458 completely sequenced and unfinished genomes—and 254 609 proteins from Swiss-Prot version 51.5 were compared, in a pairwise manner, with the program SSEARCH (http://fasta.bioch.virginia.edu/), an implementation of the SW local alignment algorithm. The sample was partitioned in blocks containing up to 2000 sequences each, and comparisons were made applying standard parameters, with an *E*-value cutoff equal to one. To overcome distortions in the *E*-value and bit score produced by the partitioning of the data, we recalculate the statistical parameters Lambda and K for each aligned pair, taking the entire dataset into account, using four different mathematical models implemented in the SSEARCH algorithm: (i) a weighted regression of average score versus library sequence length, which provides an accurate estimate of whether an alignment score is likely to occur by chance (Pearson, 1998; Pearson and Sierk, 2005), (ii) estimation from the mean and standard deviation of the library scores, without correcting for library sequence length, (iii) maximum likelihood estimates of Lambda and K and the (iv) Altschul–Gish parameters (Altschul and Gish, 1996). For each comparison, a report containing sequence identifiers, alignment length, coordinates of the most similar regions, percentage of identity, number of gaps, raw and bit scores and *E*-value was returned. These central data were subsequently connected to several third-party annotations, including gene and protein features (RefSeq), taxonomic information (NCBI Taxonomy database), gene ontology (GO), functional classification (Swiss-Prot/TrEMBL), domain and protein family classification (Pfam) and enzymatic activity (KEGG). Additionally, we have clustered all proteins of the dataset. Two or more proteins are included in the same cluster if either their SW score or the combination of identity and overlap is greater than or equal to a certain threshold (Otto *et al.*, 2008). More than 40 complete sets of clusters, using different parameter settings, were generated and stored.

The ProteinWorldDB data are stored and managed using IBM® DB2 database management system, and are publicly accessible via a web-based graphical user interface. Currently, the following analyses are implemented:

(1) Query of annotation features by primary/secondary database identifiers, GO terms, EC numbers or Pfam terms. The records are returned in tabular form, including all aforementioned qualifiers, the genome name and its NCBI taxonomy ID. This is the standard output for most results.

(2) Return of all proteins stored in the database similar to a query sequence according to a certain qualifier. The user can limit the results using the *E*-value, percentage of identity, overlap area or SW score.

(3) Comparison of protein sequences not included in the database with all proteins in the dataset using BLAST algorithm. The first five hits, including their features, are returned.

(4) Download of the complete comparison data of two (fully sequenced) genomes. The number of hits displayed can be limited as in 3.

(5) Search for unique proteins encoded by each organism. Under a given cluster threshold, these proteins represent the sequences that have not been grouped with any other sequence.

(6) Query of groups of related proteins, based on the primary/secondary database identifiers, GO terms, EC numbers or Pfam terms.

3 RESULTS

ProteinWorldDB hosts a singular core dataset, composed of nearly 4 million proteins compared in a pairwise manner with the rigorous SW algorithm, which guarantees to find a mathematically optimal alignment for a given set of parameters. With the help of the WCG, the processing took ∼7 months of calendar time (the equivalent of 3748 computer years, including an average 3-fold redundancy in the grid, which was simultaneously allocating resources to two other projects). The complete result occupies ∼1 TB in a tabular form, each line comprising 80 characters for each alignment with an *E*-value ≤1. Of the 16×10^{12} comparisons executed, 4.2×10^{9} are currently in the database (comprising 300 GB of data), corresponding to alignments with an *E*-value ≤0.001. Different groups compared subsets of sequences with a SW approach (Kanehisa *et al.*, 2006) or pairs were first filtered with a heuristic method and then compared, after satisfying a certain threshold (Rattei *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.

ACKNOWLEDGEMENTS

We wish to thank IBM®, World Community Grid™, Rede Fiocruz, Plataforma de Bioinformática PDTIS, Coordenação 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.

Boekhorst,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 genomics 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**, 63–98.

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.*, **3318**, 595–608.

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

Tian,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.