Data mining parasite genomes: haystack searching with a computer

D. LAWSON

Pathogen Sequencing Unit, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

SUMMARY

A number of genomes of parasitic organisms are presently being sequenced in the public domain, including *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei* with the likelihood of at least expressed sequence tag (EST) projects for several filarial and apicomplexan species. The early and timely release of sequence data to the community via the World Wide Web (www), and the public databases, (EMBL and GENBANK), forms an invaluable resource. Data mining, or 'haystack searching' this resource is becoming more fruitful to all members of the scientific community as the volume of data, diversity of genomes sampled, and accessibility increase.

Key words: Genomic sequence, expressed sequence tag (EST), pairwise similarity searching, high-throughput genome (HTG) sequence.

INTRODUCTION

The impact of high-throughput sequencing to the scientific community cannot be underestimated. Access to the complete genome of an organism leads the way to the compilation of the full gene complement (the transcriptome, and by analogy the proteome). The collation of such data sets has breathed new life into old technologies; namely 2D gel electrophoresis and mass spectroscopy to such an extent that individual protein spots are unambiguously identifiable using MALDI-TOF Grimm *et al*. (1998). Unique oligo-nucleotide primers can be designed for any part of the genome and sections amplified using the polymerase chain reaction (PCR). The accuracy and speed of such manipulations have opened new vistas for functional analysis, which have hitherto been complex and time consuming (for example stable transfections, gene knockouts, complementation, and expression analysis).

The past five years have seen the initiation of a number of parasite genome projects and the generation of large volumes of data. Molecular parasitology is entering a new phase in which access to the complete genome sequence for the major protozoan species is a reality. Here I discuss the ways in which the wealth of genomic sequence information can be utilised by the scientific community. I will highlight the necessary caveats of utilising preliminary data released early in sequencing process.

TYPES OF NUCLEOTIDE SEQUENCE DATA: GENOMIC V EST SEQUENCES

Nucleotide sequence data essentially comes in two types: sequences derived directly from the chromosomal DNA and sequences derived from reverse transcribed messenger RNA (mRNA).

Chromosomal DNA represents the primary data source and wherever feasible the desired end point for genome projects. The complete genomic sequence contains the coding sequences (CDSs) but also all of the control elements necessary for the transcription and translation of each and every gene. Genomic sequencing is often laborious, expensive and in some cases technically challenging (for example the extreme compositional bias of the *P. falciparum* genome). Methodologies for producing genomic sequence are essentially based on a shotgun strategy. Shotgun sequencing involves the random breaking of a DNA molecule and subsequent cloning of size-selected DNA fragments into a vector system and propagation within a bacterial host. Single colonies or plaques are picked randomly and sequenced. The scale of sequencing projects can range in size and complexity from small clone-based strategies, (*L. major*), through whole-chromosome, (*P. falciparum*), to entire genome shotguns (*Mycobacterium tuberculosis*). One important caveat to the use of genomic sequences derived from single-chromosome is that such libraries are not derived from pure material. The isolation and purification of chromosomes from pulsed-field gels will yield a greatly enriched library but not a pure one. Hence, data sets will contain clones derived from different chromosomes but the same genome. Most single chromosome libraries are in the order of 80–85% pure (Bowman, Lawson & Quail, unpublished). The cross-contamination from other chromosomes is a significantly lesser problem if dealing with assembled data sets: statistically consensus sequences of length greater than 1 kb will be derived from the selected chromosome.

Parasitology (1999) **118**, S15–S18. Printed in the United Kingdom © 1999 Cambridge University Press
https://doi.org/10.1017/S0031182099004539 Published online by Cambridge University Press

For the mRNA-derived sequences, predominantly randomly picked ESTs define the mature transcript for the coding sequence (CDS) of each gene. The construction of such libraries involves the hybridisation of a polyT+ oligonucleotide to the polyA+ tail of mRNAs and hence each clone represents a 3′ terminal portion of a gene product. It is rare to get full-length transcripts from such libraries and EST libraries give an average view of those genes expressed at the time of RNA isolation. Genes expressed in a tight temporal or spatial sense are rarely represented in randomly derived EST data sets. Putative examples of such genes include those involved in the differentiation of malarial parasites immediately after sporozoite invasion of hepatocytes, or those in a single nematode neurone. Methodologies involving multiple libraries from different life-cycle stages or extensive back hybridisation with abundant transcripts will greatly improve the rate of gene discovery (for example *Brugia malayi* and *Caenorhabditis elegans*).

QUALITY OF THE GENERATED NUCLEOTIDE DATA

Both EST and shotgun sequence reads are single pass; they represent the passage of a polymerase molecule along a DNA template and are subject to the inaccuracies and errors inherent in that process. The process of analysing the trace data from an ABI sequencing machine involves a low level of errors in the final nucleotide sequence. Hence, both data types are inherently flawed resources that may include insertion and/or deletion events (indels). The occurrence of errors becomes pronounced towards the end of the read where base calling becomes most problematic because of the non-uniform mobility of bands in gel electrophoresis.

Sequencing by a shotgun strategy depends on high redundancy coverage – that is each base is sequenced many times from independent clones. Hence, the final finished sequence represents the consensus of individual reads and is highly accurate (most genome centres work to an accuracy of 99·99 % or 1 error in 10000 bp). The generation of assembled contig consensus from the primary shotgun data is an iterative process. This data may still contain indels, but it is of a higher quality and is preferential to single shotgun reads.

Equally, most EST projects will perform cluster analysis of the full data set. Cluster analysis involves the alignment of EST reads and the generation of a set of unique consensus sequences. First, this process increases the quality of the data as each base may have been sequenced many times. Secondly, the mean sequence length increases which maximises the likelihood of identifying the transcript through similarity searches.

In terms of identifying genes the ideal data set is one of predicted CDSs, that is the primary amino acid sequence for each gene product. This reduces the size and complexity of the data set to be searched and greatly increases the sensitivity for heterologous searching. Many conserved motifs are only identifiable at the peptide level because the redundancy of the genetic code renders nucleotide comparisons inconclusive.

PARASITE GENOMES WITH HIGH-THROUGHPUT SEQUENCE DATA AVAILABLE

A number of species have been subjected to high-throughput sequencing methodologies (Table 1). Primarily these are the species highlighted by the WHO TDR program (World Health Organisation, 1999) but other species are included wherever the respective community has raised funds. Some of these have been highly successful EST projects; namely *B. malayi* and *T. gondii*. The success of parasite EST projects is reflected in their prominence in the public EST database (dBEST) (Boguski, Lowe & Tolstochev, 1993). Full genome sequencing projects are underway for *P. falciparum*, *T. brucei*, *L. major* and *T. cruzi*. These projects have already produced fruits in the completion of chromosome 2 (Gardner *et al*. 1998) from *P. falciparum*, and 1 from *L. major*, (Myler *et al*. 1999). The *T. brucei* and *L. major* projects are scaling up, promising a large volume of data in the forthcoming months. Further information is available at the respective web sites for each project. The eventual repository for all genomic and consensus EST data is the public databases (EMBL and GENBANK) in the HTG section.

Genome centres release data in an unfinished form both via their web sites and the HTG of the public databases. The early release of sequence data is vital in promoting the usage of genomic sequence data. In many cases the completion of a project may be months, if not years, away so early release is the only mechanism to allow equal access to the wealth of information available. The proper usage of the primary data can be invaluable; chromosome-specific libraries are cross-contaminated with sequences from other parts of the genome. These will not be assembled into the main project, yet are a resource for data mining to the community. Early release of data from the *P. falciparum* project has led to a number of publications based on preliminary data (Cheng *et al*. 1998; Waller *et al*. 1998).

Predicted gene transcripts are quickly incorporated into the TREMBL database, and eventually into the SWISSPROT database. TREMBL is automatically generated from EMBL flatfiles. The advantages of a rapidly updated protein database are tempered by the sometimes ambiguous annotation of

https://doi.org/10.1017/S0031182099004539 Published online by Cambridge University Press

Table 1. Parasite genomes for which high-throughput sequence is available. The availability of genomic sequence data is based on systematic sequencing of large regions of a genome.

Species	Genome size (Mb)	Genomic data	EST data
Plasmodium falciparum	30	Yes	Yes
Schistosoma mansoni	300	No	Yes
Trypanosoma brucei	25+10	Yes	Yes
Leishmania major	35	Yes	Yes
Trypanosoma cruzi	60–80	Yes	Yes
Brugia malayi	100	Yes	Yes
Onchocerca volvulus	—	No	Yes
Plasmodium vivax	30	Some	Yes
Toxoplasma gondii	80	No	Yes

the entries. SWISSPROT is a high quality database of protein sequences which is expertly annotated but by no means universal in its coverage. The rapid, almost exponential, rise in the volume of nucleotide data being produced makes it very difficult to maintain both quality and coverage.

DATA SEARCHING METHODOLOGY

The primary method of searching sequence data remains simple pairwise similarity algorithms. The traditional workhorse of similarity searching is the BLAST program (Altschul *et al*. 1990). BLAST simplifies the process of alignment by not allowing gaps (that is the output returns a series of high-scoring segment pairs or HSPs). A second generation of BLAST algorithms, BLAST2 and PSI-BLAST (Altschul *et al*. 1997), facilitate the incorporation of gaps in the alignment which can lead to alignments of higher quality. The alignment will more accurately reflect 'indel' events in nucleic acid queries and can handle polymorphic insertion regions from protein queries. The incorporation of gaps into an alignment also implies an increase in the 'noise' level, that is the number of false positives and raw BLAST results can become large and unwieldy to analyse. The problem of 'false positives' is particularly pronounced when dealing with compositionally biased genomes, for example the A+T rich DNA of *P. falciparum*, where regions of low-complexity can mask true hits.

The interpretation of gapped BLAST output requires some form of moderation, whether that be formalised in a script or via human inspection. The trade-off between speed and accuracy of simple pairwise alignments supplied by these algorithms comes into consideration when dealing with the large data sets that high-throughput methodology can provide. Essentially, when searching for gene homologues human inspection will allow appraisal on a gene by gene basis. Appraisal of large data sets such as EST consensus sequences mapped to underlying genomic DNA is best achieved using scripts. The moderation of the raw BLAST output at a number of stringency levels will reduce the data set into those sequences definitely mapped, those not mapped and a third set that lies in a 'grey area' of similarity score. It is this set that then requires human appraisal to identify the reason for the ambiguous similarity data. Matches may be due to poor quality sequence, the compositional bias of the genome or the degree of repetitive sequences present in the genome.

QUERY CHARACTERISTICS

Similarity searching is the equivalent of '*in silico*' hybridisation experimentation – not a stunning insight into the nature of the search but a useful metaphor for the process. The parameters utilized by the algorithm and subsequent analysis of the results are primarily dependent on the query characteristics, whether the query is derived from homologous or heterologous sequences. For example, searches involving an EST read against the underlying species genomic database (BLASTN) have the expectation of a 100% identical match, taking into account 'indel' events toward the 3′ of the EST query. The appraisal of such analysis is straightforward and uncomplicated. On the other hand, searches utilising heterologous queries, for instance one peptide sequence from species A against the genomic database of species B (TBLASTN), are unlikely to give 100% identity results.

The implication of such similarity searching is to identify sequences that have diverged from a shared 'ancestral' sequence. Utilising similarity searching as a method of identifying genes is an inherently flawed process. The highest-scoring alignment statistically is not necessarily the most parsimonious structural one. Equally, convergent evolution can mean that two genes fulfil the same role in the cell but share very little sequence similarity. A large number of bacterial genomes have been sequenced to date and analysis suggests that approximately 25–35% of predicted genes are unsupported hypotheticals; that is there is no significant similarity to any known gene. This figure highlights the shortcomings of similarity searching as a method of gene identification. The appraisal of similarity data between heterologous species requires a greater understanding of the nature of genetic drift, the taxonomic relationship between species and an aspect of the domain structure of proteins.

DISCUSSION

Molecular parasitology is rapidly entering the genomic era where access to complete genomes is a reality. The timely release of unfinished data is important to drive research in communities which

https://doi.org/10.1017/S0031182099004539 Published online by Cambridge University Press

have previously undertaken 'piecemeal' sequencing of small regions of DNA. Sequence data are available in a number of different formats, genomic and cDNA, but both are essentially focused on gene discovery. Analysis of these data resources is best achieved using web-based tools, especially pairwise similarity searching algorithms, and such tools are becoming more powerful and readily usable by the whole community. The rapid advances made elsewhere in genomic research ensure that our understanding of parasite genomes and, by inference, parasite biology will be greatly augmented in the coming years.

REFERENCES

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., MYERS, E. W. & LIPMAN, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

BOGUSKI, M. S., LOWE, T. M. & TOLSTOSHEV, C. M. (1993). dbEST–database for 'expressed sequence tags'. *Nature Genetics* **4**, 332–333.

CHENG, Q., CLOONAN, N., FISCHER, K., THOMPSON, J., WAINE, W., LANZER, M. & SAUL, A. (1998). *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Molecular and Biochemical Parasitology* **97**, 161–176.

GARDNER, M. J., TETTELIN, H., CARUCCI, D. J., CUMMINGS, L. M., ARAVIND, L., KOONIN, E. V., SHALLOM, S., MASON, T., YU, K., FUSII, C., PEDERSON, J., SHEN, K., JING, J., ASTON, C., LAI, Z., SCHWARTZ, D. C., PERTEA, M., SALZBERG, S., ZHOU, L., SUTTON, G., CLAYTON, R., WHITE, O., SMITH, H. O., FRASER, C. M., ADAMS, M. D., VENTER, J. C. & HOFFMAN, S. L. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132.

GRIMM, R., GRASSER, K. D., KUBACH, J. & HANCOCK, W. S. (1998). A rapid and sensitive procedure for the micro-purification and subsequent characterization of peptides and protein samples by N-terminal sequencing and matrix assisted laser desorption ionization time of flight mass spectrometry. *Journal of Pharmacology and Biomedical Analysis* **18**, 545–554.

MYLER, P. J., AUDLEMAN, L., DE VOS, T., HIXSON, G., KISER, P., LEMLEY, C., MAGNESS, C., RICKEL, E., SISK, E., SUNKIN, S., SWARTZELL, S., WESTLAKE, T., BASTIEN, P., FU, G., IVENS, A. & STUART, K. (1999). *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proceedings of National Academy of Sciences, USA* **96**, 2902–2906.

WALLER, R. F., KEELING, P. J., DONALD, R. G. K., STRIEPEN, B., HANDMAN, E., LANG-UNNASCH, N., COWMAN, A., BESRA, G. S., ROOS, D. S. & MCFADDEN, G. I. (1998). Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proceedings of National Academy of Sciences, USA* **95**, 12352–12357.

WORLD HEALTH ORGANISATION (1999). TDR program WWW home page. *http://www.who.int/tdr/*.

https://doi.org/10.1017/S0031182099004539 Published online by Cambridge University Press