FEBS Letters 451 (1999) 354–355 FEBS 22059

Genomics

Reproducibility in genome sequence annotation: the *Plasmodium falciparum* chromosome 2 case

Sophia Tsoka[a], Vasilis Promponas[a,b], Christos A. Ouzounis[a,*]

[a] *Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK*
[b] *Department of Biology, University of Athens, Athens, Greece*

Received 9 April 1999

Key words: Genomics; Sequence analysis; *Plasmodium falciparum*

With the recent publication of the complete sequence of the *Plasmodium falciparum* chromosome 2, an additional step towards the understanding of the biology of this important human pathogen has been made [1]. Hopefully, the genome sequence will provide a basis for further study and the possible identification of targets for drug and vaccine development. This project follows the model of eukaryotic genome sequencing, one complete chromosomal sequence at a time, which was initiated with the yeast chromosome III [2,3]. Progress has been made by developing systems that perform automated genome sequence annotation and can further accelerate the experimental analysis [4]. Yet, computational genomics still depend on the insight of the analysis teams and loosely used terminology, despite the fact that biological research requires strict protocols and method specifications.

1. The question of reproducibility

A reason for this discrepancy comes from the fact that the computational analysis of genome sequences is still largely an imprecise process and occasionally non-reproducible [5]. Although automated genome sequence annotation systems offer reproducibility of results and clear specifications of the criteria employed [6], they have been regarded as inadequate for some purposes [7]. The end result today is a vast collection of genome sequence annotations by various groups (for 20 complete, and many partial, genomes at the time of writing[1]) that may be conflicting or inconsistent and usually impossible to compare.

By its very nature, the crucial step of genome annotation should be carried out with the utmost precision and clarity [8]. Any genome sequence annotation project should be reproducible and well-documented. To achieve this, accurate procedures and protocols as well as adequate representational devices are required [9]. Until this goal is reached, it is compelling that genome annotation groups should be extremely cautious in making their results clear, publicly available and, at the very least, reproducible.

2. Comparing notes

Currently, there is a general lack of comparative studies for different annotation projects. Although it is well-known (and expected) that differences may exist, there is no quantification of the degree of conflict between various analysis projects. One exception is the genome of *Methanococcus jannaschii*, which is being continually annotated over a period of 3 years[2]. That comparison showed that there can be a level of conflict as high as 10% between different groups and methodological approaches [10]. To examine the recent results for the *P. falciparum* chromosome 2, we have performed an analysis of the 210 gene sequences, employing methods identical to those reported in the original publication[3] [1].

3. The *P. falciparum* chromosome 2

Our efforts have yielded strikingly different results for the *P. falciparum* chromosome 2 sequence. It is surprising that despite the virtual absence of false positives, only 124 out of 210 (59%) cases are in general agreement (Fig. 1). The ambition for high coverage in the original analysis appears both to compromise clarity and render a comparison very difficult [1]. We list below some major categories of cases where annotation can be misleading and we identify the sources of conflict and provide some recommendations.

(1) We have observed seven cases where the presence of a domain was used for the annotation of the entire sequence. For instance, the C-terminal domain of PFB0520w displays similarities to serine/threonine kinases, while its N-terminal domain remains unique. This open reading frame has been described as a 'novel prt kinase' by the original authors [1]. Functional descriptions should only be accepted if homology covers the full length of the query sequence. Otherwise, domain similarity should be clearly stated in the annotation

*Corresponding author. Fax: (44) (1223) 494471.
E-mail: ouzounis@ebi.ac.uk

[1] http://geta.life.uiuc.edu/–nikos/genomes.html, mirrorred at: http://www.ebi.ac.uk/research/cgg/genomes.html

[2] http://geta.life.uiuc.edu/–nikos/MJannotations.html, mirrorred at: http://www.ebi.ac.uk/research/cgg/annotation/MJannotations.html.

[3] BLAST(p) version 1.4.8, 2.0 and psi-BLAST against the non-redundant protein sequence database (nrdb) at the EBI (version 23 February 1999 with 374 547 sequences and 111 115 040 residues): a new method of compositional bias masking called CAST ([6], Promponas et al., in preparation) was used. Only similarity-based prediction methods were employed in the present study.

PII: S0014-5793(99)00599-2

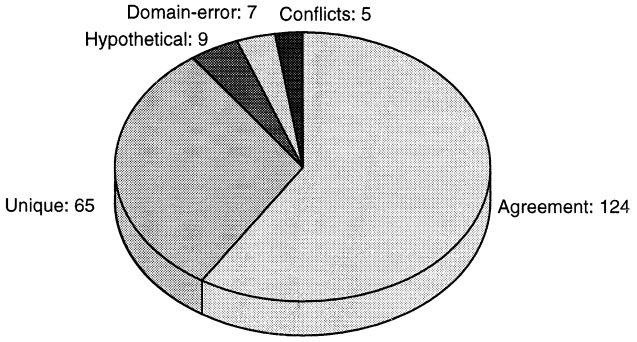

Fig. 1. Distribution of agreement for the sequence annotations of the *P. falciparum* chromosome 2 sequence. Of the 210 total gene sequences, 124 were in agreement between the original and the current analysis. There are 65 proteins that are unique in the database but not annotated as such, nine proteins that were not characterised as 'hypothetical' (with homologues of unknown function), seven proteins that were characterised on the basis of the presence of a domain only and finally, five cases that cannot be classified in any of the above categories. A table containing detailed information is available at ⟨http://www.ebi.ac.uk/research/cgg/annotation/pf2.html⟩.

(a working definition here could be that a domain is present when the similar sequence covers less than half the length of the query sequence).

(2) It is important to unambiguously distinguish similarity-based from 'ab initio' predictions as well as to provide the source of annotation for all query sequences characterised by similarity. Examples are PFB0125c and PFB0980w, both of which have been originally described as 'predicted membrane-associated prt' [1]: PFB0125c is unique in the database, while PFB0980w is similar to PFB0085c, PFB0920w and PFB0925w (which in turn have been correctly characterised as 'prt with DnaJ domain (RESA-like)' [1]). Two additional problems with the 'ab initio' predictions are that the accuracy of the programs employed is not discussed and the terms used ('OO', 'TM', 'membrane-associated' versus 'integral membrane prt') are not explicitly defined [1]. These annotations are incorporated into the public databases and become a potential source of error propagation. We have seen nine cases that are 'predicted' to be membrane/secreted proteins or enzymes (Fig. 1).

(3) For the cases where functional prediction is not possible, it is necessary to specify whether they have no detectable homologues in the database (with a specified cut-off score and pairwise sequence comparison methods, defined as unique) or exhibit similarities to proteins of unknown function (usually described as hypothetical). This simple naming convention has been proven useful in distinguishing between species-specific and ubiquitous genes [4]. The authors report that 43% of the genes have no detectable homologues, but this number is virtually unobtainable. Some non-similarity-based predictions can also be found in this category (for example, PFB0780c which we report as a hypothetical protein with similarity to PFB0770c, has been originally described as 'predicted integral membrane prt' [1]). In total, we have observed 65 cases that have not been properly annotated as 'unique' sequences in the database (Fig. 1).

(4) Finally, we were unable to trace the source of the assignment for five cases. These can be considered as non-reproducible. Although some of these cases may be false positives, it is not possible to classify them in any of the above mentioned categories and they can only be taken as conflicting annotations. One example is PFB0535w (similar to vanadate resistance proteins, originally characterised as 'predicted multiple TM membrane prt', without an indication for the source of this assignment [1]).

Even if the 65 cases mentioned under case (3) are considered as agreeing in principle, the 21 remaining cases would represent exactly 10% of the total number of genes analysed. It is worth asking whether it is of any value to risk over-predictions and/or non-reproducible results for such a number of additional predictions or refrain from any annotation to avoid such conflicts.

4. Concluding recommendations

When results are not readily reproducible, it may be wise to neglect terms such as 'hypothetical', 'predicted' or other, very poor, predictions altogether. An alternative is to use these terms in a consistent manner, as is done in SWISS-PROT (Rolf Apweiler, personal communication). As more computational biology groups collaborate with primary genome sequencing centres, the strategies for the deposition of the data should be reconsidered. In the past, the definitions of most sequence records in the database were derived from experimental analysis, while the same fields today are used for computational predictions.

This comparative study underlines the subjective nature of the sequence annotation process. It is imperative that annotators strive for high precision, re-use of similar terms, clarity of definitions and reproducibility of results [11]. A conservative approach to genome sequence annotation may indeed form a first step towards specifications of restricted vocabularies for the description of the molecular function.

Acknowledgements: The authors would like to thank Rolf Apweiler (EBI/SWISS-PROT), Liisa Holm (EBI), Peter Karp (Pangea Systems) and Nikos Kyrpides (University of Illinois) for useful comments. This work was supported by the European Molecular Biology Laboratory and TMR Grant ERBFMRXCT960019 from the EU DGXII (Science, Research and Development).

References

[1] Gardner, M.J., Tettelin, H., Carucci, D.J., Cummings, L.M., Aravind, L., Koonin, E.V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J. and Shen, K. et al. (1998) Science 282, 1126–1132.
[2] Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, I., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P. and Benit, P. et al. (1992) Nature 357, 38–46.
[3] Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) Nature 358, 287.
[4] Casari, G., Andrade, M.A., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, J., Valencia, A. and Sander, C. (1995) Nature 376, 647–648.
[5] Kyrpides, N.C. and Ouzounis, C.A. (1999) Mol. Microbiol. 32 (in press).
[6] Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. and Sander, C. (1999) Bioinformatics 15, (in press).
[7] Bork, P. and Koonin, E.V. (1998) Nat. Genet. 18, 313–318.
[8] Kyrpides, N.C. and Ouzounis, C.A. (1998) Science 281, 1457–1457.
[9] Karp, P.D. (1998) Bioinformatics 14, 753–754.
[10] Andrade, M., Casari, G., de Daruvar, A., Sander, C., Schneider, R., Tamames, J., Valencia, A. and Ouzounis, C. (1997) Comput. Appl. Biosci. 13, 481–483.
[11] Fleischmann, W., Möller, S., Gateau, A. and Apweiler, R. (1999) Bioinformatics 15, (in press).