BIOINFORMATICS

Vol. 19 no. 6 2003, pages 717–726
DOI: 10.1093/bioinformatics/btg077

Evaluation of annotation strategies using an entire genome sequence

Ioannis Iliopoulos[1,†], Sophia Tsoka[1], Miguel A. Andrade[2,3], Anton J. Enright[1,‡], Mark Carroll[1,§], Patrick Poullet[1,¶], Vassilis Promponas[4], Theodore Liakopoulos[4], Giorgos Palaios[4], Claude Pasquier[4,‖], Stavros Hamodrakas[4], Javier Tamames[5,**], Asutosh T. Yagnik[5,††], Anna Tramontano[5,‡‡], Damien Devos[6], Christian Blaschke[6], Alfonso Valencia[6], David Brett[3,§§], David Martin[7,¶¶], Christophe Leroy[8], Isidore Rigoutsos[9], Chris Sander[8,‡] and Christos A. Ouzounis[1,*]

[1]Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK, [2]EMBL Heidelberg, D-69012 Heidelberg, Germany, [3]Max-Delbrück Centre, D-13122 Berlin, Germany, [4]Department of Cell Biology and Biophysics, University of Athens, GR-15701 Athens, Greece, [5]Deparment of Computational Biology, IRBM, I-00040 Rome, Italy, [6]CNB-CSIC, Campus University Autonoma, E-28049 Madrid, Spain, [7]Biotechnology Centre, University of Oslo, N-0349 Oslo, Norway, [8]Whitehead Institute, MIT, Cambridge, MA 02139, USA and [9]IBM Research Center, Yorktown Heights, NY 10598, USA

Received on May 22, 2002; revised on October 29, 2002; accepted on November 22, 2002

ABSTRACT

Motivation: Genome-wide functional annotation either by manual or automatic means has raised considerable concerns regarding the accuracy of assignments and the reproducibility of methodologies. In addition, a performance evaluation of automated systems that attempt to tackle sequence analyses rapidly and reproducibly is generally missing. In order to quantify the accuracy and reproducibility of function assignments on a genome-wide scale, we have re-annotated the entire genome sequence of *Chlamydia trachomatis* (serovar D), in a collaborative manner.

Results: We have encoded all annotations in a structured format to allow further comparison and data exchange and have used a scale that records the different levels of potential annotation errors according to their propensity to propagate in the database due to transitive function assignments. We conclude that genome annotation may entail a considerable amount of errors, ranging from simple typographical errors to complex sequence analysis problems. The most surprising result of this comparative study is that automatic systems might perform as well as the teams of experts annotating genome sequences.

Availability and supplementary information: http://www.ebi.ac.uk/research/cgg/annotation/cteval/.

Contact: ouzounis@ebi.ac.uk

INTRODUCTION

While the amount of genome sequence information increases exponentially, the annotation of genome sequences appears to be lagging behind both in terms of quality and quantity. Several researchers have been involved in a significant improvement of the annotations provided by the original genome sequencing groups, both providing annotations for previously uncharacterized genes and also correcting a number of errors due to false similarity detection. Several such cases have been reported, for example for the first three genomes that were

*To whom correspondence should be addressed.
† INA-EKETA, GR-57001 Thessaloniki, Greece
‡ Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA
§ Aetion Technologies LLC, Worthington, OH 43085, USA
¶ Institut Curie, F-75248 Paris, France
‖ CNRS, UMR6543, F-06108 Nice, France
** Alma Bioinformatics, E-28760 Madrid, Spain
†† Cap Gemini Ernst & Young, London SW1X 7LX, UK
‡‡ Univ. of Rome 'La Sapienza', I-00185 Rome, Italy
§§ MWG-Biotech AG, Ebersberg, D-85560 Berlin, Germany
¶¶ Wellcome Trust Biocentre, Univ. of Dundee, Dundee DD1 5HN, UK

completely sequenced, namely *Haemophilus influenzae* (Casari *et al.*, 1995), *Mycoplasma genitalium* (Ouzounis *et al.*, 1996), and *Methanococcus jannaschii* (Andrade *et al.*, 1997; Koonin *et al.*, 1997; Kyrpides *et al.*, 1996a). It has been noted that occasionally annotations may not be reproducible (Brenner, 1999; Kyrpides and Ouzounis, 1999; Tsoka *et al.*, 1999), because no sufficient evidence for certain predictions is made available, rendering systematic comparisons of results difficult.

Problems with genome annotation include inconsistent function descriptions, false (positive or negative) assignments, unsupported predictions, haphazard use of various terms. In addition, the absence of highly reliable sets that could be used as a 'gold standard' to benchmark methods or annotation strategies and the general lack of structured terminology classifications (ontologies) pose significant challenges to comparative studies of annotation by computational means.

This generally unsatisfactory situation arises from a number of factors. First, the range of known molecular functions is mostly based on a few well-characterized species (e.g. *Escherichia coli*) or systems (e.g. tryptophan biosynthesis Crawford, 1989). This results in a large number of extrapolations on the basis of sequence similarity. Second, all known functions derive from a limited number of proteins from the database and there is significant recycling of terms and definitions, usually without a trace of their origin. Third, potential annotation errors may propagate and result in falsely characterized cases which can 'infect' more recent sequence database entries.

Automatic genome annotation has been approached with the development of automatic systems such as PEDANT (Frishman *et al.*, 2001) and GeneQuiz (Andrade *et al.*, 1999). These systems can be used as tools that accelerate the task of human experts by providing detailed and exhaustive information for function assignments. Moreover, GeneQuiz attempts to reproduce the manual steps involved in genome sequence annotation (Andrade *et al.*, 1999). The ultimate goal of such projects is the fully automatic annotation of large sequence collections with a performance similar to that of human experts. To achieve this goal, however, it is imperative to understand the quality of annotation through comparative studies.

To address some of the above issues, we evaluated the quality of annotations for the complete genome of *Chlamydia trachomatis* (serovar D) (Stephens *et al.*, 1998). We have asked the following questions: First, how good were the original annotations and were they reproducible? Second, if the same query genome was analysed against the same databases automatically, would there be an improvement over the original analysis? Third, if we repeated the same analysis manually, how would our results compare to the original (also manually derived) or the subsequent (automatically derived) sets of annotations? Despite the importance of annotation comparisons, there has been little work carried out so far, usually in the form of claims for 'improved' function predictions from various groups against the original work. Herein, we address the above questions and focus on the identification of the potential sources of problems both for manual and automatic annotation approaches.

METHODS

Data analysis

The *Chlamydia trachomatis* (serovar D) genome (893 ORFs) was obtained from ⟨http://chlamydia-www.berkeley.edu:4231/⟩ (Stephens *et al.*, 1998). The database used for similarity searches both for automatic and manual analyses was the nrdb database at the EBI (29 Jan 1999 version, 372 471 sequences). Any novel findings are based on that database, and we have refrained from using later releases, so that the comparison with the original work is as meaningful as possible. The GeneQuiz system was used on a 8-CPU SGI Challenge. Total CPU time was approximately 40 hours.

Manual annotation

This analysis has been performed in the context of a collaborative network of nine laboratories. The genome sequence was divided and distributed between all groups. The evaluation was carried out over the internet (using both e-mail and www access) and one meeting. The overhead of communication was significant but results were cross-checked between all participating laboratories and the manual analysis was repeated at least three times by different people.

During the manual analysis, iterative PSI-BLAST (Altschul *et al.*, 1997) searches were performed with default parameters and up to five iterations per sequence. All query sequences were filtered for compositionally biased regions using CAST (Promponas *et al.*, 2000) and default parameters. Great care was taken to admit function assignments only from database protein entries with experimentally determined functions, by manually accessing the publication records of those entries. In the future, a database (or a section of existing databases) where experimentally determined functions are tagged would be an extremely valuable resource for computational genome analysis (Karp, 1998).

To further validate the manually obtained annotations, all ORFs were also annotated using a recently developed, exhaustive annotation system (Rigoutsos *et al.*, 2001), based on the Bio-Dictionary, a collection of patterns that are generated by processing very large sequence databases (Rigoutsos *et al.*, 1999), using the TEIRESIAS algorithm (Rigoutsos and Floratos, 1998). The rationale behind using the Bio-Dictionary system was that an annotation

approach which is fundamentally distinct from the use of similarity searches would: (a) provide independent corroboration of manually derived annotations and (b) potentially annotate sequences whose weak similarities to characterized proteins are not detectable by BLAST (Floratos *et al.*, 2001).

All results were summarized and validated, using the Genome AnnoTatiOn System (GATOS) (Karp and Ouzounis, in preparation)[†]. All analyses including the original dataset at the time of publication, the automatic analysis (dated 12 December 1998) and the 'final' manual evaluation are available at : http://www.ebi.ac.uk/research/cgg/annotation/cteval/.

Encoding scheme

We have realized that, in order to make all annotation sets comparable, we had to encode the available information into a structured format. We have used the original annotations as described, the automatically derived annotations directly from the GeneQuiz database and we have encoded our own manual annotations using the GATOS system. Other standards for genome annotation, such as DAS (Dowell *et al.*, 2001), were not considered suitable because they concentrate mostly on gene structure. Below, we provide some explanations regarding the encoding conventions employed.

The GATOS system uses six principal object classes to encode species, genetic element (e.g. chromosomes), gene, product, protein family and protein complex. The most critical object class in this encoding scheme for the representation of protein function is the 'product' class. We have only used the 'product' class for the encoding of function assignments.

For protein function representation, four fields are used: 'enzyme', 'similar-to', 'function' and 'domain'. The first and second types are always accompanied by at least one Enzyme Commission (EC) number, indicating the precise biochemical reaction type that is believed to be catalyzed by the corresponding enzyme. 'Similar-to' is a simple convention that permits the recording of potential function without committing to the enzyme specificity. 'Function' is a general field that applies to all other biochemical roles, except enzymes. Finally, 'domain' allows the user to record the presence of a protein domain or motif function. For the purposes of this study, we consider all cases of genes containing identified domains as additional assignments (possibly of unknown function).

Multi-functional enzymes are recorded by simply using sequential annotations (multiple fields per record) to capture this property, without an explicit reference to sequence positions. For Enzyme Commission (EC) numbers, we have used the 'X.Y.Z.-' instead of the 'X.Y.Z.99' convention, to document unclassified biochemical reactions for enzymes. No terms such as 'putative', 'probable' or 'possible' were used for our assignments.

In naming components of multi-subunit complexes, we have followed naming conventions from the literature, where possible. We have used the word 'subunit' 52 times and the word 'chain' 18 times. The latter includes all DNA polymerase components (usually called 'chains'), as well as five other enzymes (tryptophan synthase, ribonucleoside-disphosphate reductase, riboflavin synthase, phenylalanyl-tRNA synthetase and succinyl-CoA synthetase). We have not attempted to annotate protein complexes for *Chlamydia trachomatis* in GATOS.

During the analysis, if the query sequence was longer than the target sequence from which function was inferred, we have recorded this fact (either in the 'similar-to' field or noted as a domain). If the query sequence was shorter, we have accepted the functional assignment only if the alignment corresponded to a segment of the target sequence with characterized function.

For practical purposes and in order to assess typographical mistakes, we use original annotations *verbatim* by enclosing them in <brackets>, e.g. <phopholipase> (it should read 'phoSpholipase').

Scoring scheme

All 893 annotations from the original and the automatic datasets were compared against our 'final' annotation set. For each comparison, a score was manually recorded according to the Transitive Annotation-Based Scale (Table 1)—or TABS (Ouzounis and Karp, 2002). This scale describes the major cases of errors in genome annotation and ranks them according to their effects on error propagation in the database. TABS is a qualitative scale, therefore we have not attempted to generate numerical results, such as means, medians or total sums. We have found that it covers most, if not all, of the possible sources of annotation conflicts. In the TABS system, underpredictions and false negatives are not as heavily penalized as overpredictions and false positives, because they are less likely to propagate. Domain errors are an intermediate situation, considered to be undesirable and are heavily penalized because of their potential to propagate flawed assignments (Smith and Zhang, 1997).

Before we proceed further, we will describe in detail the actual conventions that we have employed in our scoring scheme. First, we scored each case only once, with the maximum possible penalty. For example, if there was a domain error and a typographical error, the penalty would be 5 (domain error) (Table 1). It is envisaged that more complex scoring schemes may be implemented in the future. Second, all 'tentative' assignments from

[†] Jointly developed by Peter D. Karp (SRI International) and Christos A. Ouzounis (EMBL-EBI)—see also: http://www.ebi.ac.uk/research/cgg/ontology/gatos/.

Table 1. The Transitive Annotation-Based Scale (TABS) used in this analysis

Score	Description
7	False positive
6	Over-prediction
5	Domain error
4	False negative
3	Under-prediction
2	Undefined source
1	Typographical error
0	Total agreement

Table 2. The evaluation of the original and the automated annotation

Score	Original	Genequiz	Clear	Tentative	Marginal	Unknown
7	43	35	28	7	0	0
6	50	62	60	2	0	0
5	20	23	21	2	0	0
4	26	50	10	0	30	10
3	67	48	46	2	0	0
2	84	64	60	4	0	0
1	35	10	10	0	0	0
0	565	598	326	6	172	94
No score	3	3	2	0	1	0
Total	893	893	563	23	203	104

Column names reflect the two different analyses (see Methods); the breakdown of the assignment levels for GeneQuiz is also shown. Three genes were not scored, due to recent publications (see Methods).

GeneQuiz (Andrade *et al.*, 1999) were considered as genuine assignments, whereas all 'marginal' assignments were considered as non-assignments (i.e. corresponding to 'hypothetical' proteins). This important distinction has an effect on scoring, for instance: a tentative wrong assignment would qualify as false, while a marginal wrong assignment would not (depending on the final annotation). Third, wherever protein domains (of known or unknown function) were detected, the corresponding entries were considered to be characterized, in order to achieve some consistency with other cases (e.g. CT256, a CBS domain-containing protein). If the original or automatic annotations did not describe the presence of the domain, this was considered as an underprediction.

To exemplify our scoring strategy, we describe some illustrative cases in some detail. Typographical errors may seem benign and indeed they are not penalized heavily with the TABS system. However, we have found cases such as CT369 <dehyroquinate synthase> and CT370 <shikimate 5-dehyrogenase> (D is missing in both cases: dehyDro-). Interestingly, both SWALL and MedLine contained 28 entries each with the term 'dehyrogenase' as of 20 September 2001. Unfortunately, even simple typographical errors can lead to seriously flawed analyses and interpretations (Kyrpides and Ouzounis, 1998). In this particular case, for instance, many of the typographical errors we have observed have already propagated to the genome annotations of other *Chlamydia* species or strains (e.g. CT485, reported as <ferrochetalase> also in *Chlamydia pneumoniae* (Kalman *et al.*, 1999)). Penalties for undefined source of information were given to unclear assignments such as leader peptides, transmembrane segments and all other instances of non-homology based predictions.

In case multiple disagreements are detected, cases are scored only once, with the highest corresponding value of TABS. Such examples are: (i) CT742, considered to be an overprediction (RNA methyltransferase and not <rRNA MethyltransferAse>)—(A is missing); (ii) CT673, also considered as an overprediction, although there is a domain error as well; (iii) CT595, considered as a domain problem, although there is a typo as well (thioL:disulfide interchange protein) (see web site for details).

We have not scored three cases where more recently published information has resulted in annotation updates in the corresponding database entries, which may have not been available to the original authors. These cases are CT580 (Grass *et al.*, 2000), CT771 (Thorne *et al.*, 1995) and CT804 (Lange and Croteau, 1999; Rohdich *et al.*, 2000).

RESULTS

Comparison to the original annotation

The original analysis has generated 604 function assignments (representing 68% of total) (Stephens *et al.*, 1998). We have analysed these data and identified a number of errors, including typographical errors that make any computational access of these data impossible. Following the Transitive Annotation-Based Scale (Table 1) we have scored all original annotations according to the degree of seriousness for error propagation in the databases. There are only 565 cases (out of 893—or 63%, including no functional assignments, i.e. <hypothetical protein>), with a total agreement (score = 0) to the final annotation (Table 2).

We have also found 35 typographical errors (score = 1), some of which have already been transitively assigned to other database entries (e.g. CT084 characterized as <Phopholipase D Superfamily> and CT485 described as <Ferrochetalase>, while the correct terms should have been phospholipase and ferrochelatase, respectively—see also Table 3). Other typographical errors include assignments of various kinds, including minor inconsistencies (e.g. CT044 and CT341) (Table 3).

Table 3. Examples of inconsistent annotations

Gene-ID	Tabs	Original annotation	Final annotation
CT044	1	<SS DNA Binding Protein>	Single-stranded DNA-binding protein SSB
CT084	1	<Phopholipase D Superfamily>	Endonuclease nuc homolog
CT186	1	<Glucose-6-P Dehyrogenase>	Glucose-6-phosphate dehydrogenase
CT329	1	<Exodoribonuclease VII>	Exonuclease VII large subunit
CT341	1	<Heat Shock Protein J>	DnaJ protein
CT369	1	<Dehyroquinate Synthase>	3-dehydroquinate synthase
CT451	1	<Phosphatidate Cytidylytransferase>	Phosphatidate cytidylyltransferase
CT485	1	<Ferrochetalase>	Protoheme ferro-lyase
CT586	1	<Exinuclease ABC Subunit B>	Excinuclease ABC subunit B
CT073	2	<Predicted OMP>	–
CT131	2	<(Possible Transmembrane Protein)>	–
CT175	2	<Oligopeptide binding protein permease>	Oligopeptide binding protein OppA
CT323	2	<Initiation Factor IF-1>	Translation initiation factor IF-1
CT422	2	<Possible metalloenzyme>	–
CT546	2	<Predicted OMP>	–
CT655	2	<KDO Synthetase>	KDO-8-phosphate synthetase
CT709	2	<Rod Shape Protein-Sugar Kinase>	Rod shape-determining protein MreB
CT857	2	<[IM protein]>	Na+/H+ antiporter NhaD homolog
CT069	3	<Integral Membrane Protein>	ABC-3 integral membrane ATPase TroC
CT106	3	<Predicted pseudouridine synthetase family>	Ribosomal large subunit pseudouridine synthase C
CT403	3	<rRNA Methylase (SpoU family)>	tRNA (Gm18) 2′-O-methyltransferase
CT732	3	<Ribityllumazine Synthase>	Riboflavin synthase beta chain
CT370	5	<Shikimate 5-Dehyrogenase>	3-dehydroquinate dehydratase / Shikimate 5-dehydrogenase
CT555	5	<SWI/SNF family helicase>	snf2/rad54 family helicase C-terminus
CT613	5	<Dihydropteroate Synthase>	Dihydropteroate synthase FolP /2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase FolK
CT708	5	<SWI/SNF family helicase>	snf2/rad54 family helicase C-terminus
CT295	6	<Phosphomannomutase>	Phosphohexomutase
CT381	6	<Arginine Binding Protein>	Extracellular solute-binding protein
CT637	6	<Aromatic AA Aminotransferase>	Amino-acid aminotransferase class I
CT799	6	<General Stress Protein>	Ribosomal protein L25
CT815	6	<Phosphomannomutase>	Phosphohexomutase
CT844	6	<Cytosine deaminase>	Cyclic amidine deaminase
CT046	7	<Histone-like protein 2>	KARP Chlamydial protein
CT068	7	<rRNA methylase>	ABC transporter, ATPase TroB
CT074	7	<ABC superfamily ATPase>	RecF protein
CT141	7	<Protein Translocase>	–
CT312	7	<Predicted ferredoxin>	–
CT320	7	<Transcriptional termination protein>	Transcription antitermination protein NusG
CT402	7	<ATPase>	Lipid A 4′-kinase
CT454	7	<Arginyl tRNA Transferase>	Arginyl-tRNA Synthetase
CT840	7	<PP-loop superfamily ATPase>	-

Table is sorted by gene identifier and TABS category. Original and final annotations are also shown. The GATOS functional assignment categories are omitted from the final annotation, for clarity (available on the web site).

In the undefined source category (score = 2), we have identified 84 cases (Table 2), most arising from a loose usage of terms and other types of inaccuracies. For instance, CT073 and CT546 are characterized as <predicted OMP>, an assignment that contains a term not defined either in the original paper or the corresponding web site. Other examples are given in Table 3, along with our final annotations. All results are provided on the above mentioned web site.

In the category of underpredictions (score = 3), we have counted 67 cases (Table 2). These cases represent assignments where the set of methods used can reveal some more specific biochemical functions than the ones reported in the original publication. Examples are provided in Table 3.

We have identified 26 false negative cases (score = 4), which are listed in Table 4. Some of the most interesting assignments here correspond to genes CT071 (a reductoisomerase involved in terpenoid biosynthesis) (Takahashi *et al.*, 1998), CT257 (a CBS domain-containing protein), CT356 (a thioredoxin domain-containing protein), CT359 (a biotin synthesis protein BioY homolog), CT473 (an alpha-hemolysin homolog), CT627 (a rhodanese domain-containing protein), CT650 (RecA), CT700 (a TPR domain-containing protein) and CT718 (Flagellar assembly protein FliH) (Table 4). In addition, we report a second RsbU-like protein (CT589), adjacent to the one that has been detected originally (CT588) (Stephens *et al.*, 1998).

Domain errors (score = 5) were 20 and include CT370 (original annotation: <Shikimate 5-Dehyrogenase>—note a typographical error), CT555 and CT708 both having N-terminal domains not common with snf2/rad54 family helicase proteins (Stephens *et al.*, 1998) and CT613 (a bifunctional protein containing the corresponding enzymes encoded by bacterial genes FolP and FolK) (original annotation corresponds only to the FolP enzyme) (Table 3).

Overpredictions (score = 6) were 50 and include CT295 and CT815 (Phosphohexomutase EC 5.4.2.-), CT637 (a class I aminotransferase homolog EC 2.6.1.-), CT799 (ribosomal protein L25) and CT844 (a cyclic amidine deaminase EC 3.5.4.-) (Table 3).

Finally, the most severe penalty was imposed to false positive assignments, which in our opinion cannot be supported by evidence present in the public databases (score = 7). We detected 43 such cases, with highly conflicting assignments, which include CT068 (original annotation: <rRNA methylase>, final annotation: ABC transporter, ATPase TroB), CT074 (original annotation: <ABC superfamily ATPase>, final annotation: RecF protein), CT320 (original annotation: <Transcriptional termination protein>; final annotation: Transcription antitermination protein NusG, containing a KOW domain (Kyrpides *et al.*, 1996b)), CT402 (original annotation: <ATPase>; final annotation: Lipid A 4'-kinase EC 2.7.1.130), CT454 (original annotation: <Arginyl tRNA Transferase> which represents a totally different biochemical function (Kwon *et al.*, 1999), final annotation: Arginyl-tRNA Synthetase EC 6.1.1.19) (Table 3).

Overall, the categories 1–4 contain functional assignments that are innocuous as far as error propagation in databases is concerned. In total, the original annotations that were classified in these categories amount to 212 instances (out of 893, or 24% of total) (Table 2). Total agreement, recorded in category 0, represents 63% of the 893 genome entries (Table 2). The other three categories 5–7 (domain errors, overpredictions and false positive cases) contain 113 instances (13% of total), according to our analysis. These are the cases with the highest probability of error propagation in the public databases.

Comparison to the automatic annotation

The automatic analysis by GeneQuiz resulted in 563 'clear' and 23 'tentative' function assignments (Andrade *et al.*, 1999) (representing 66% of total). The remaining 307 genes are split between 203 'marginal' assignments (considered as cases of undetected homologies) and 104 'unknown' genes (of which 73 have homologues of unknown function and 31 genes without homologues in the database) (Table 2). We have analysed these data in exactly the same way as the original annotation using TABS (Table 1).

As described in Methods, 'marginal' cases from the automated analysis were considered as 'hypothetical' and were penalized as false negatives if a function was manually assigned by us (30 cases), while the ones without a final function were not penalized (172 + 1 cases) (Table 2). It is interesting that of the 23 'tentative' cases, almost half of them are overpredicted (domain error, overprediction or false positive).

There are only 598 cases (out of 893 or 67%, including no functional assignments), with a total agreement (score = 0) to the final annotation (Table 2). It is interesting to note that in most cases there has been considerable agreement with the final annotation, within the 'semantic boundaries' of function assignment, i.e. within a range of variations of function description names or synonyms. It is also notable that only 454 cases were considered to be in agreement for all annotation (both scores were set to zero), representing only a fraction of the genome information (51%). This set of annotations, however, independently confirmed by three different analyses, should be one of the most reliable function assignment collections that can be used for the benchmarking of genome annotation strategies.

A significant problem with the automatic analysis is the issue of near-circular annotations, i.e. function assignments essentially deriving from the query sequence, possibly via transitive annotation to close homologues. This is mainly due to the insufficient discriminatory capacity of these systems to distinguish the native query sequence from highly similar ones (or itself, after low-complexity masking (Promponas *et al.*, 2000)). In fact, this problem has appeared more often for well-annotated families with many members. If query sequences were not included in the databases, this problem would have been solved. However, automatic systems such as GeneQuiz can reliably transfer annotation, even for quite complex cases (e.g. CT557—dihydrolipoamide dehydrogenase EC 1.8.1.4).

Since the system derives all annotations from the databases without human intervention, it is not pertinent

Table 4. Novel findings in this analysis, considered as false negatives in the original paper (Stephens *et al.*, 1998)

Gene-ID	Original annotation	*	Final annotation
CT057	Hypothetical protein	F	GcpE protein
CT071	Hypothetical protein	E	1-deoxy-D-xylulose 5-phosphate reductoisomerase EC -.-.-.-
CT077	Hypothetical protein	F	Thiamine biosynthesis lipoprotein ApbE precursor
CT166	Hypothetical protein	F	Toxin B N-terminus
CT255	Hypothetical protein	F	Beta-lactamase regulatory homolog MazG
CT257	Hypothetical protein	D	CBS
CT274	Hypothetical protein	F/D	Secretion chaperone SscB homolog TPR
CT277	Hypothetical protein	F/D	Proprotein cleavage endoprotease furin homolog furin
CT287	PP-loop superfamily ATPase	E	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase EC 2.1.1.61
CT339	Hypothetical protein	F	Competence protein ComEC/Rec2 homolog
CT356	Hypothetical protein	D	Thioredoxin
CT357	Hypothetical protein	F/F	Chelated iron transport system membrane protein ABC transporter, ATP-binding protein
CT359	Hypothetical protein	F	Biotin synthesis protein BioY
CT450	YaeS family (length 253)	E	Undecaprenyl pyrophosphate synthetase EC 2.5.1.31
CT456	Hypothetical protein	D	NYDD-unknown
CT473	Hypothetical protein	F	Alpha-hemolysin homolog
CT589	Hypothetical protein	F	C-terminus PP2C serine phosphatase RsbU homolog
CT606	Hypothetical protein	S	NTPase HAM1 homolog EC 3.6.1.15
CT610	Hypothetical protein	F	Coenzyme PQQ synthesis protein C
CT627	Hypothetical protein	D	Rhodanese
CT650	Hypothetical protein	F	RecA protein
CT691	Hypothetical protein	F	Phosphate transport system protein PhoU
CT700	Hypothetical protein	D	TPR
CT718	Hypothetical protein	F	Flagellar assembly protein FliH
CT736	Hypothetical protein	F	Phosphatidylethanolamine binding protein homolog
CT741	Hypothetical protein	F	Preprotein translocase complex subunit YajC

* Functional assignment categories in GATOS: F: function; E: enzyme; S: similar-to; D: domain. Gene-IDs in *italics* represent the five gene assignments which GeneQuiz has also identified and in full agreement with the final annotation.

to discuss in detail minor errors due to typographical mistakes in database records or parsing problems (for a full breakdown, see Table 2). We have marked 10 typographical errors (score = 1) (e.g. CT061 predicted as 'RNA POLYMERASE SIGMA FACT') and 64 assignments from undefined source (score = 2) (e.g. CT740 predicted as 'PHENOLHYDROXYLASE COMPONENT' on the basis of its similarity to protein HI0171 from *Haemophilus influenzae*, TrEMBL accession number O05012). There were also 48 under-predictions (score = 3) (e.g. CT410 predicted as 'PROBABLE POLY(A) POLYMERASE (EC 2.7.7.19) (PAP)'; final annotation: 'Poly(A) polymerase EC 2.7.7.19') and 50 false negatives (score = 4) (e.g. CT633 with no assignment by GeneQuiz, annotated as porphobilinogen synthase EC 4.2.1.24). The system cannot yet cope with multiple assignments due to the multi-domain structure of proteins, therefore there were 23 cases of domain error (score = 5) (e.g. CT370—see above and Table 3 for the correct annotation). Overpredictions (score = 6) amounted to a total of 62 cases (e.g. CT258 characterized as 'TRNA SPLICING PROTEIN SPL1' instead of a homologue of the NifS-like Class V aminotransferases (Ouzounis and Sander1993)). Finally, false positives (score = 7) were 35 in total (e.g. CT141—see above and Table 3).

It is instructive to observe the breakdown of the GeneQuiz predictions, on the basis of the categories of confidence level: 'clear' (BLAST p-value $< 10^{-10}$), 'tentative' (p-value $< 10^{-04}$), 'marginal' (p-value $< 10^{-01}$) and 'unknown' (p-value $> 10^{-01}$) (Andrade *et al.*, 1999) (Table 2). There are two patterns emerging from this breakdown: first, the total number of 'clear' (563) and 'unknown' (104) cases represent the most confident predictions by the system (667 out of 893 cases—or 75%) but only 492 of them (74%) have a score less than 3; second, the majority of 'tentative' and 'marginal' cases (183 out of 226—or 81%) have a score less than 3 and are considered as acceptable annotations (Table 2). It is thus evident that most of the mistakes arise from the 'clear' assignments that suffer from over- or under-prediction (Table 2). A precision estimate for the GeneQuiz system, given these strict criteria, can be taken as the 492/667 ratio, or 74% of the most reliable set of predictions.

Remarks on the final annotation

Our manual analysis resulted in 586 function assignments (representing 66% of total). Although we have no way

to test the validity of our predictions, these are based on detailed manual analysis and evaluation of the GeneQuiz runs, the identification of weak similarities and domain patterns as well as iterative searches against the database (see Methods).

It is encouraging that annotation using the Bio-Dictionary approach resulted in 862 cases (out of 893, or 96.5%) of total agreement with our manual annotations. Of the remaining 31 cases, 13 were annotated manually but were not detected by the Bio-Dictionary, whereas 18 cases were annotated by the Bio-Dictionary but missed by BLAST and, consequently, by the manual annotation (details are available on the web site).

We have made every effort to use a consistent terminology for our assignments, re-use terms for the description of paralogous proteins, list all functions using accepted database description lines and finally make all results available on our web site. One potential omission is the total absence of descriptions such as 'hypothetical' or 'unique' protein. In the past, we have proposed that proteins with homologues but no known functions may be called 'hypothetical' while unique proteins in the database (i.e. proteins with no homologues in the database) should be listed as such (Tsoka *et al.*, 1999). Since this distinction can be carried out independently by mere sequence comparison and the string 'hypothetical' does not convey any function information, we decided to drop this convention in this present project.

Comparison between the original and the automatic methods

Due to space limitations, we cannot discuss the pairwise combinations of penalty levels for the original and automatic methods (see Table 2)—the results are available on the web site for further analysis.

Our results show that: (i) overall, the comparison of three strategies results in considerable agreement (Figure 1); (ii) the particular original annotation contained a variety of significant errors or omissions (Tables 3 and 4); (iii) the performance of automatic annotation is comparable to the original annotation (Table 2); and (iv) automatic systems may actually suffer mostly from erroneous database description lines and less from false-positive homology detection.

Finally, we believe that the set of 454 cases scored for total agreement between these approaches could serve as a 'gold standard' that might be used for benchmarking purposes, especially for complex cases of low sequence similarity detection, such as the presence of short motifs.

DISCUSSION

The challenge in the post-genomics era is to characterize a multitude of biochemical functions in a genome-wide context, through such approaches as the identification of gene clusters participating in the same cellular processes, as well as the indication of biochemical role for genes of unknown function (Eisenberg *et al.*, 2000). As the information incorporated into databases is increasing and being refined over time, there will be more opportunities to generate more reliable predictions and more complex hypotheses. However, it should be stressed that the success of such a scenario implies that database information captures knowledge on molecular function in a specific (no false positives) and a sensitive manner (no false negatives).

With the advent of completely sequenced genomes (Kyrpides, 1999), the amount of sequences that can be used for annotation is increasing exponentially. The load of genome sequence annotation may be alleviated by 'analysis robots', such as GeneQuiz (Andrade *et al.*, 1999), programs that embody key rules as applied by human experts and use them at great speed in a fully automated fashion. The power of such systems lies not only in the increased speed of analysis, allowing updates that take advantage of the latest information in databases, but also in the fact that the analysis is consistent and therefore results are comparable across different species and time points (Iliopoulos *et al.*, 2000). However, their success is compromised by: (i) incomplete or misleading database annotations and (ii) the lack of suitable benchmarking procedures.

Concerning the transfer of function from a previously characterized (experimentally or computationally) molecule to a query sequence, there needs to exist a balance between a conservative approach to annotation (risking underprediction) and a more aggressive approach

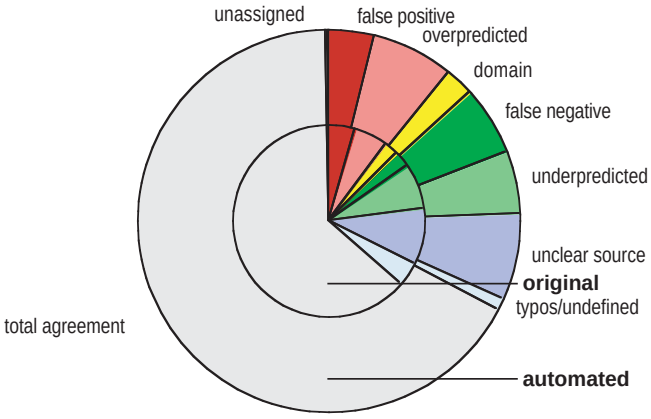

Fig. 1. Comparison of the original and automatically derived approaches for the annotation of the *Chlamydia trachomatis* genome. The eight categories of annotation errors from the Transitive Annotation-Based Scale (TABS) (Table 1) are shown. The corresponding counts for each category can be found in Table 2.

(risking overprediction). The former case relates to cases where partial transfer of biochemical information occurs (e.g. failing to detect short motifs which carry functionally important signals, complex domain architectures for multi-functional proteins or substrate specificity). In the latter case, assignments may be made without supporting evidence (e.g. detection of false similarities or functionally divergent proteins). Evidently, tipping the balance towards a more conservative approach may be preferable, as overprediction results in error propagation (Karp, 1998).

In this study, we have analyzed a relatively large set of hundreds of annotations, have compared the performance of different approaches and have identified potential sources of error. The original annotation appears to be of high quality, as few false negatives were identified (Table 4). Similar results were obtained from the automated annotation set, which is encouraging, perhaps with the exception of false negatives which are twice as many for the automatic approach. However, the agreement of only 454 cases between these two sets (51%) presents a challenge and may be due to the different sources of errors from human experts or computer algorithms.

To address the two points mentioned above, namely assessing the performance of automated sequence analysis systems and evaluating the effect of erroneous database annotations, it is suggested that consistent benchmarking mechanisms are set up, whereby quality of annotation is evaluated, sources of errors are traced and 'guidelines' to be set are determined accordingly. In the absence of formal benchmarking test-suites, such as the ones used in the protein structure prediction field (Venclovas *et al.*, 1999), systematic evaluations of functional sequence analyses, such as the one presented here, are of significant importance.

Guidelines

Some guidelines, stemming from the present detailed analysis, are summarized below, in order of increasing complexity:

- Eliminate or reduce false positives by conservative assignment of function, to avoid error propagation in the databases.
- Indicate clearly whether the assignment is based on experimental or computational analysis. Since experimental evidence is generally more reliable, such a reference would provide an indication of level of confidence for the specific assignment.
- Distinguish the *ab initio* from the similarity-based assignments for all cases of computational prediction. In the first case (e.g. transmembrane predictions), the method applied should be mentioned along with an evaluation of its performance (Tsoka *et al.*, 1999). In the case of function assignment based on similarity, it is important to cite the source of assignment, in order to facilitate reproducibility.
- Use restricted vocabularies for function description, when possible. Even though currently it is difficult to impose strict rules on definitions of function, workers should try to re-use terms for proteins of similar function (Ashburner *et al.*, 2000).
- Define strict protocols for function assignment. These protocols should also specify procedures of error correction for information already deposited in the databases. Curated databases should also document with precision their internal protocols of quality control and updates.
- Provide annotations in a highly structured, computationally accessible form to facilitate data exchange, comparison and analysis.
- Enhance the social nature of the genome sequence annotation process. We have experienced an acute lack of sophisticated collaborative environments that would augment efficiency, compared to the available internet technologies. The work was always carried out in pairs, which we termed the 'buddy' system for annotation. We have also found that re-annotation is probably significantly more difficult than a first-pass annotation, since erroneous results can impede performance.

It is important to realize that there is no such thing as an 'easy case' in genome sequence annotation.

ACKNOWLEDGEMENTS

This work was fully supported by the Training and Mobility for Researchers (TMR) Programme of the European Commission (DG-XII Science, Research and Development). We also thank the British Council and the General Secretariat for Research Technology, Greece for additional travel support.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,M., Casari,G., de Daruvar,A., Sander,C., Schneider,R. *et al.* (1997) Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Comput. Appl. Biosci.*, **13**, 481–483.

Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.

Casari,G., Andrade,M.A., Bork,P., Boyle,J., Daruvar,A. *et al.* (1995) Challenging times for bioinformatics. *Nature*, **376**, 647–648.

Crawford,I.P. (1989) Evolution of a biosynthetic pathway: the tryptophan paradigm. *Annu. Rev. Microbiol.*, **43**, 567–600.

Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.

Floratos,A., Rigoutsos,I., Parida,L. and Gao,Y. (2001) DELPHI: a pattern-based method for detecting sequence similarity. *IBM J. Res. Dev.*, **45**.

Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A. *et al.* (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.

Grass,G., Grosse,C. and Nies,D.H. (2000) Regulation of the *cnr* cobalt and nickel resistance determinant from *Ralstonia* sp. strain CH34. *J. Bacteriol.*, **182**, 1390–1398.

Iliopoulos,I., Tsoka,S., Andrade,M.A., Janssen,P., Audit,B. *et al.* (2000) Genome sequences and great expectations. *Genome Biol.*, **2**, interactions0001.0001–0001.0003.

Kalman,S., Mitchell,W., Marathe,R., Lammel,C., Fan,J. *et al.* (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.*, **21**, 385–389.

Karp,P.D. (1998) What we do not know about sequence analysis and sequence databases [editorial]. *Bioinformatics*, **14**, 753–754.

Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.*, **25**, 619–637.

Kwon,Y.T., Kashina,A.S. and Varshavsky,A. (1999) Alternative splicing results in differential expression, activity, and localization of the two forms of arginyl-tRNA-protein transferase, a component of the N-end rule pathway. *Mol. Cell Biol.*, **19**, 182–193.

Kyrpides,N.C. (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.

Kyrpides,N.C., Olsen,G.J., Klenk,H.-P., White,O. and Woese,C.R. (1996a) *Methanococcus jannaschi* genome: revisited. *Microb. Comp. Genomics*, **1**, 329–338.

Kyrpides,N.C. and Ouzounis,C.A. (1998) Errors in genome reviews. *Science*, **281**, 1457–1457.

Kyrpides,N.C. and Ouzounis,C.A. (1999) Whole-genome sequence annotation: 'going wrong with confidence'. *Mol. Microbiol.*, **32**, 886–887.

Kyrpides,N.C., Woese,C.R. and Ouzounis,C.A. (1996b) KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. *Trends Biochem Sci.*, **21**, 425–426.

Lange,B.M. and Croteau,R. (1999) Isopentenyl diphosphate biosynthesis via a mevalonate-independent pathway: isopentenyl monophosphate kinase catalyzes the terminal enzymatic step. *Proc. Natl Acad. Sci. USA*, **96**, 13714–13719.

Ouzounis,C., Casari,G., Valencia,A. and Sander,C. (1996) Novelties from the complete genome of *Mycoplasma genitalium*. *Mol. Microbiol.*, **20**, 898–900.

Ouzounis,C. and Sander,C. (1993) Homology of the NifS family of proteins to a new class of pyridoxal phosphate-dependent enzymes. *FEBS Lett.*, **322**, 159–164.

Ouzounis,C.A. and Karp,P.D. (2002) The future of genome-wide re-annotation. *Genome Biol.*, comment2001.2001–2001.2006

Promponas,V.J., Enright,A.J., Tsoka,S., Kreil,D.P., Leroy,C. *et al.* (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Rigoutsos,I., Floratos,A., Ouzounis,C., Gao,Y. and Parida,L. (1999) Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. *Proteins*, **37**, 264–277.

Rigoutsos,I., Huynh,T., Floratos,A., Parida,L. and Platt,D. (2001) Dictionary-driven protein annotation. IBM TJ Watson Research Center: #TR-RC22262.

Rohdich,F., Wungsintaweekul,J., Luttgen,H., Fischer,M., Eisenreich,W. *et al.* (2000) Biosynthesis of terpenoids: 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase from tomato. *Proc. Natl Acad. Sci. USA*, **97**, 8251–8256.

Smith,T.F. and Zhang,X. (1997) The challenges of genome sequence annotation or 'the devil is in the details'. *Nat. Biotechnol.*, **15**, 1222–1223.

Stephens,R.S., Kalman,S., Lammel,C., Fan,J., Marathe,R. *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, **282**, 754–759.

Takahashi,S., Kuzuyama,T., Watanabe,H. and Seto,H. (1998) A 1-deoxy-D-xylulose 5-phosphate reductoisomerase catalyzing the formation of 2-C-methyl-D-erythritol 4-phosphate in an alternative nonmevalonate pathway for terpenoid biosynthesis. *Proc. Natl Acad. Sci. USA*, **95**, 9879–9884.

Thorne,N.M., Hankin,S., Wilkinson,M.C., Nunez,C., Barraclough,R. *et al.* (1995) Human diadenosine 5′,5‴-P1,P4-tetraphosphate pyrophosphohydrolase is a member of the MutT family of nucleotide pyrophosphatases. *Biochem. J.*, **311**, 717–721.

Tsoka,S., Promponas,V. and Ouzounis,C.A. (1999) Reproducibility in genome sequence annotation: the *Plasmodium falciparum* chromosome 2 case. *FEBS Lett.*, **451**, 354–355.

Venclovas,C., Zemla,A., Fidelis,K. and Moult,J. (1999) Some measures of comparative performance in the three CASPs. *Proteins*, (suppl.), 231–237.