Opinion

# The past, present and future of genome-wide re-annotation

Christos A Ouzounis* and Peter D Karp†

Addresses: *Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK. E-mail: ouzounis@ebi.ac.uk †Bioinformatics Research Group, AI Center, SRI International, Menlo Park, CA 94025, USA. E-mail: pkarp@ai.sri.com

Published: 31 January 2002

*Genome* ***Biology*** 2002, **3(2)**:comment2001.1–2001.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/2/comment/2001



## Abstract

Annotation, the process by which structural or functional information is inferred for genes or proteins, is crucial for obtaining value from genome sequences. We define the process of annotating a previously annotated genome sequence as 're-annotation', and examine the strengths and weaknesses of current manual and automatic genome-wide re-annotation approaches.

> *It is perhaps hard to make firm statements on such questions without having examined them many times*
> Aristotle, *Categories*, 8b21 (translated by J.L. Ackrill, Clarendon Press, Oxford 1963)

## Why re-annotate?

Over the past ten years, we have witnessed the publication of several chromosomes or complete genome sequences from a variety of bacterial, archaeal and eukaryotic species. The trend towards genome sequencing is expected to continue or even accelerate in the foreseeable future. The wealth of sequence information being produced has generated the need for rapid annotation and subsequent biological interpretation of genome sequences. Annotation can be defined as a process by which structural or functional information is inferred for genes or proteins, usually on the basis of similarity to previously characterized sequences in public databases. The annotation process associates genome sequences with functional information and guides experimentation by relating genotypes to phenotypic properties.

Once a genome-sequencing project is completed and the information is released into the public domain, it is common practice for certain groups of researchers to take a 'second look' at the original annotation, for various reasons. We define the process of annotating a previously annotated genome sequence as 're-annotation'. Motivations for re-annotation include discovery of more genes and protein functions, testing and performance-comparison of existing or newly developed annotation methods, and assessment of annotation reproducibility. Re-annotation also provides up-to-date information for end-users, using the latest resources - such as new or improved algorithms and richer databases.

Clearly, the drive for re-annotation goes back in time, arising even before the availability of entire genome sequences. For example, in an attempt to assign function to a number of uncharacterized, hypothetical genes from archaeal species, one of the earliest large-scale re-annotation studies produced a number of novel predictions [1]. What sets whole-genome re-annotation apart from other analyses, however, is its distinctly discontinuous and comparative nature: independent groups of researchers systematically generate novel predictions and compare them with an original set of gene-function predictions in an incremental and stepwise manner. We believe that this mode of annotation provides valuable insights into the process of protein-function assignment. It is different from the continuous mode of annotation adopted by the groups who originally annotate genome sequences having completed them, as they may not always document their annotation methods, or improvements to annotations, in the published record.

Genome-wide re-annotation is characterized by a number of distinct elements. The groups who perform re-annotation usually have no access to the original primary sequencing data (such as fluorescence traces from genome-sequencing machines), making the detection of certain features - such as

frameshift errors - quite cumbersome. The process is labor-intensive, because it considers the entire genome and involves a significant number of manual operations, including the correction of misleading original annotations. Finally, the lack of 'gold standards' by which to judge annotations represents a challenge and raises a serious, but by no means unique, reproducibility issue: because there is no 'right' answer for annotation, how can we assess its success? Quality control of annotation is probably the most important technical aspect, because it provides critical information on the performance of various approaches, by correcting various errors (higher precision) or generating more predictions (higher coverage); these issues are considered further below. Re-annotation has not attracted sufficient attention as a distinct and specialized subject, possibly because of the highly charged and competitive nature of genome bioinformatics [2]. Here, we examine as objectively as possible the strengths and weaknesses of current re-annotation approaches and suggest a number of possible improvements.

## Re-annotated genomes

Despite progress in the field of computational genomics, the process of annotation is still a largely manual, labor-intensive endeavor [3]. Because of the large number of genome sequence entries currently available (over 300,000 genes), no single group has ever been able to generate manual annotations for all proteins. Yet there is a great need for up-to-date, exhaustively annotated genome sequences. Thus, systems such as GeneQuiz [4], which infers sequence annotations automatically, provide valuable information resources; for instance, we have recently been able to generate 73,500 gene annotations for 31 sequenced genomes [5]. During the year 2001, another 30 genomes have been re-analyzed (P.J. Janssen and C.A.O., unpublished observations), and their annotations can be accessed online [6].

Re-annotation projects for individual species have been reported in the literature by a handful of groups. The species re-annotated include (with strain names omitted for brevity): *Haemophilus influenzae* [7-10], *Mycoplasma genitalium* [10-12], *Methanococcus jannaschii* [13-16], various archaeal species [17], *Mycoplasma pneumoniae* [18], *Chlamydia trachomatis* [15], *Thermotoga maritima* [19], *Saccharomyces cerevisiae* [20-24], *Plasmodium falciparum* (chromosome II) [25], *Aeropyrum pernix* [26], and isolated cases of single genes [27] (Table 1). One interesting, and encouraging, pattern to emerge from these studies is that the level of improvement provided by re-annotation, calculated by expressing the number of genes for which new functions are predicted as a percentage of the total number of genes in the genome, is on average 7% (Table 1). This indicates that, for the most part, various groups using different methods generate sets of predictions that are generally quite similar. These percentages can also be considered to represent the level of disagreement between the various groups (as a function of genome size). Proteins can be classified into two broad categories: assigned to a predicted function or unassigned (sometimes referred to as 'hypothetical'); the improvement rate usually refers to the re-assignment of hypothetical proteins to proteins of predicted function.

## Measures of annotation accuracy

The above mentioned re-annotation reports usually claim that there has been an improvement over the original (or previous) attempts for genome annotation. But in all these predictions it is not certain whether this improvement represents more accurate identification of a gene or a protein function that escaped detection from the previous analysis. Indeed, improvements over previous under-predictions (or false negatives) may correspond to current over-predictions (or false positives). There is always pressure when re-annotating a genome to produce a 'better' result, which can easily be obtained by loosening the criteria for function prediction - for example by using a weaker threshold for sequence-similarity comparisons. A cautionary note is therefore appropriate here: when researchers embark on a re-annotation project, the expectation is that they will be able to assign more functions to a set of sequences using computational methods. This natural tendency is usually supported either by looser thresholds in the analysis or by more up-to-date (but not necessarily richer) supporting databases. Thus, the 'better' results may be questionable, because of the subjectivity associated with any manual analysis. We believe the real challenge is therefore to procure and implement objective standards for genome annotation quality.

Given a gold standard - a completely correct set of annotations - two measures of accuracy can be defined. First, coverage is defined as the ratio of true positives over the sum of true positives plus false negatives - so, if there are no false negatives, coverage is 100%. Second, precision is defined as the ratio of true positives over the sum of true positives plus false positives - so, if there are no false positives, precision is 100%. In any analysis, there is a trade-off between coverage and precision. A combined measure of these two numbers is accuracy, which is defined as the ratio of true (positive plus negative) cases over the total number of cases (where 'cases' are, for example, the number of genes or proteins). Although these measures have not always been used explicitly in genome-annotation projects, they are usually implied in arguments about prediction accuracy.

The problem is that even if we agree on measures of coverage, precision and accuracy, we do not currently have a single gold standard for genome-wide annotation. There is no complete genome for which all the gene structures (start sites, exons, introns, and so on) and the encoded protein functions have been experimentally determined. Thus, all annotation attempts remain tentative - especially in the case of protein function assignments and descriptions. To assess

**Table 1**

**Re-annotation projects**

| What | How | Why | New/Total | % | Who | When | Reference |
|---|---|---|---|---|---|---|---|
| Yeast chromosome III | The very first re-annotation; 17 new predictions compared to the original 57 (74 total assignments) | C, E | 17/182 | 9.3 | Bork *et al.* | 1992 | [20,21] |
| Yeast chromosome III | Subsequent re-annotation; 19 predictions over the 74 above (93 total assignments) | C | 19/171 | 11.1 | Koonin *et al.* | 1994 | [22] |
| Various archaeal species | One of the first large-scale analyses, but not genome-wide | E | 30/95 | 31.6* | Ouzounis *et al.* | 1995 | [1] |
| Yeast chromosome VIII | Re-annotation | C | 24/269 | 8.9 | Ouzounis *et al.* | 1995 | [24] |
| *Haemophilus influenzae* | Automated genome annotation; 148 new assignments over previous 1,007 (1,155 total assignments) | C, E | 148/1,743 | 8.5 | Casari *et al.* | 1995 | [9] |
| *Haemophilus influenzae* | Additional gene findings | A | 17/1,743 | 0.1 | Robison *et al.* | 1996 | [10] |
| *Haemophilus influenzae* | Re-annotation and metabolic reconstruction; 1,408 total assignments (cf. 1,155 above) | E | 253/1,703 | 14.9 | Tatusov *et al.* | 1996 | [8] |
| *Haemophilus influenzae* | Metabolic reconstruction | E | Individual cases | N/A | Karp *et al.* | 1996 | [7] |
| *Mycoplasma genitalium* | Additional gene findings | A | 3/470 | 0.6 | Robison *et al.* | 1996 | [10] |
| *Mycoplasma genitalium* | Re-annotation | C | 21/470 | 4.5 | Ouzounis *et al.* | 1996 | [11] |
| *Methanococcus jannaschii* | Manual re-annotation | C | 214/1,738 | 12.3 | Kyrpides *et al.* | 1996 | [14] |
| *Methanococcus jannaschii* | Re-annotation; reproducibility study | C, F | 23/1,682 | 1.4 | Andrade *et al.* | 1997 | [13] |
| *Saccharomyces cerevisiae* | Short open reading frame identification | A | 10/6,357 | 0.2 | Andrade *et al.* | 1997 | [23] |
| Various species | Cautionary statement | B | Individual cases | N/A | Smith and Zhang | 1997 | [3] |
| *Methanococcus jannaschii* | Cautionary statement | B | Individual cases | N/A | Kyrpides and Ouzounis | 1998 | [34] |
| *Methanococcus jannaschii* | Cautionary statement | B, D | 20/1,738 | 1.2 | Kyrpides and Ouzounis | 1999 | [15] |
| *Chlamydia trachomatis* | Cautionary statement | B, D | 10/893 | 1.1 | Kyrpides and Ouzounis | 1999 | [15] |
| *Campylobacter jejuni* | Cautionary statement | B | Individual cases | N/A | Pallen *et al.* | 1999 | [27] |
| *Methanococcus jannaschii* | Additional gene findings | A | 31/1,773 | 1.8* | Raghavan and Ouzounis | 1999 | [17] |
| *Methanobacterium thermoautotrophicum* | Additional gene findings | A | 13/1,871 | 0.7* | Raghavan and Ouzounis | 1999 | [17] |
| *Archaeoglobus fulgidus* | Additional gene findings | A | 27/2,409 | 1.1* | Raghavan and Ouzounis | 1999 | [17] |
| *Pyrococcus horikoshii* | Additional gene findings | A | 42/2,061 | 2.0* | Raghavan and Ouzounis | 1999 | [17] |
| *Plasmodium falciparum* chromosome II | Re-annotation; reproducibility study | F | 21/210 | 10.0 | Tsoka *et al.* | 1999 | [25] |
| *Mycoplasma genitalium* | Comparison of other annotations, reproducibility study | F | Individual cases | N/A | Brenner | 1999 | [33] |
| *Aeropyrum pernix* | COGs matching | E | 315/2,694 | 11.7 | Natale *et al.* | 2000 | [26] |
| *Pyrococcus abyssii* | COGs matching | E | Individual cases | N/A | Natale *et al.* | 2000 | [26] |
| *Mycoplasma genitalium* | Contextual analysis | E | 21/480 | 4.4 | Huynen *et al.* | 2000 | [12] |
| *Mycoplasma pneumoniae* | Contextual analysis plus experiments | E | 109/688 | 15.8 | Dandekar *et al.* | 2000 | [18] |
| *Thermotoga maritima* | Contextual analysis | E | 193/1,877 | 10.3 | Kyrpides *et al.* | 2000 | [19] |
| **In total: 10+ species** | | | | **Key result: 7 ± 5% (*excluded)** | **Approximately ten groups** | **Nine years** | **23 papers** |

Column names and explanations: What, species or chromosome; How, comments or methods; Why, the reasons for re-annotation - A, to find more genes; B, a cautionary statement; C, to find more functions; D, to achieve fewer errors; E, using new methods; F, to assess reproducibility; % , the improvement - in terms of additional genes predicted - over previous annotations, as a percentage of the total number of genes in the genome; Who, authors; When, publication year; Reference, citation. N/A denotes not applicable. * Denotes percentages that have not been taken into account for the calculation.

the reproducibility of sequence-similarity-based annotation, some studies have focused on the set of known enzymes and used the Enzyme Commission (EC) classification number to measure performance: it has been shown in this way that function prediction by homology never reaches 100% accuracy [28,29]. Comparison of genome annotations with proteins of known structure also suggests that 100% accuracy is unattainable, although this result may also be influenced by the smaller number of structures than sequences in the public databases [30].

In short, what we have learned in the past few years can be of only relative, not absolute, value. We have been able to compare various different approaches to genome annotation, but we are ultimately not aware of the absolute accuracy of these predictions. Only experimental verification of gene and protein assignments will conclusively address the performance of computationally based function assignments on a genome-wide scale. This prospect may be within reach in the near future, thanks to the initiatives of structural and functional genomics, but at present we should be seeking a working definition of annotation accuracy for comparison purposes. The absolute 'truth' for annotation quality could be defined as the maximum amount of information a panel of experts can generate on the basis of computational analysis, which extends beyond standard homology-based prediction and takes into account metabolic pathway analysis [31], contextual function prediction [12] and whole-genome analysis and comparison [32]. This level of annotation should reflect the maximum amount of information that can currently be generated for a given genome sequence, without the inclusion of errors that may propagate [33,34].

## Quantifying annotation quality

Unfortunately, the above definition of annotation accuracy still contains a subjective factor, which corresponds to the 'panel of experts'. Obviously, this can be very controversial, because it strongly depends on the composition of such a hypothetical panel and the opinions of the panel members. Indeed, one interesting experiment would be to assess the degree to which different expert predictions agree by using a set of blind predictions over identical genome datasets, similar in scope and spirit to the CASP competition for predicting protein structures [35]. One important issue here is the need for continuous tracking of experimentally verified gene or protein functions. No matter how accurate computational predictions become, all annotations should in principle be traceable to biochemical or genetic experiments that derived a function for a gene product in the first place. This is an issue that has not been sufficiently addressed in the current databases, and there is an acute lack of annotation source and history for a large number of protein database entries, making this task exceedingly difficult. Moreover, there is no good mechanism in place for automatically monitoring biological experiments that are pertinent to a particular system under consideration and integrating this information with computational analysis, for example by using databases of published literature [36].

In our own re-annotation projects, we have recently come up with a qualitative scale of genome annotation quality, called the transitive annotation-based score, or TABS (Table 2). This score is based on a number of criteria, as follows. First, it represents a distance scale between two annotation attempts. For example, if during re-annotation a particular annotation is considered to have been a false positive, a high penalty is assigned. This does not necessarily imply that the previous annotation was an error but rather that the distance between the two attempts is very high. Second, penalty scores are ranked according to their potential damaging effects when propagated in the databases. For example, an over-prediction is potentially more detrimental than an under-prediction, because all homologs are in danger of

**Table 2**

**Transitive annotation-based scale (TABS): a qualitative distance scale for the assessment of annotation reproducibility in genome projects**

| Score | Description | Comment |
|---|---|---|
| 7 | False positive | Original annotation predicts function without any supporting evidence |
| 6 | Over-prediction | Original annotation predicts a specific biochemical function without sufficient supporting evidence |
| 5 | Domain error | Original annotation overlooks different domain structure of query and reference proteins |
| 4 | False negative | Original annotation does not provide predicted function although there is sufficient evidence to characterize the query protein |
| 3 | Under-prediction | Original annotation predicts a nonspecific biochemical function although a more detailed prediction could have been made |
| 2 | Undefined source | Original annotation contains undefined terms, non-homology based predictions, and so on |
| 1 | Typographical error | Original annotation contains typographical errors that may be propagated in the database |
| 0 | Total agreement | Original annotation is correct, but annotations may be only semantically (but not computationally) identical |

Column names and explanations: Score, the score between two assignments; Description, a description of the potential disagreement between two projects; Comment, explanatory comments for ranking/scores. We consider scores of 0-3 as relatively benign compared to scores of 4-7, as the latter have a much more significant impact on genome sequence and database quality.

inheriting this assignment, usually without a trace. Third, we opted for a mutually exclusive sum of penalties. For instance, if we detect both a domain error and a typographical error, we penalize the case as having a domain error, because we consider it much more important (Table 2). The TABS scale for annotation quality departs from the traditional 'percentage' notion of successful assignment and, we believe, provides a first step towards more quantifiable comparisons; the cumulative sum of assignments may be considered as a function of the error-propagation potential of the individual assignments. In the future, more complex schemes may be devised that would also weigh individual assignments according to the confidence of the prediction. For example, given two equivalent annotation schemes and assuming no mistakes, the one with the higher confidence should score higher - but errors in high-confidence predictions should result in high penalties.

## Best practice, and open questions

It is fortunate that institutions such as The Institute for Genomic Research (TIGR) [37], which has sequenced so many genomes to date, have opted for a high-precision approach. The fact remains, however, that many portions of available genome sequences have yet to be annotated, because researchers hold back from inferring functions with low certainty. Occasionally, people who work with particular species have found new homologies that are informative of function, but these results have not been incorporated into any existing database. Clearly, there needs to be a robust mechanism by which this occurs - and in the meantime we can only urge all researchers to make their results available in central databases. There are serious attempts to build consensus genome annotations for a number of model organisms; these projects are community-based initiatives to keep annotations current and of high quality [38]. Unfortunately, this is not the case for less well-studied organisms with smaller research communities. We would argue that one obvious way is to provide annotations for all genomes that have been sequenced using automatic, reproducible protocols. Such protocols have been developed in projects such as GeneQuiz [4] and PEDANT [39]. We have recently analyzed all genomes available up to year 2001 using GeneQuiz - these can be accessed online [6]. Automatically generated annotations may not be carefully curated but they do provide a solid basis, upon which the community can build. Frequently, such annotations may be imported into curated databases, such as, for example, SWISS-PROT [40].

Another, more complex, issue is the peer-reviewing process for genome-sequencing articles. Results in the published papers that result from sequenced genomes are thoroughly reviewed by expert referees - but the only experimental result of these papers is the genome sequence of the species under consideration. All annotations represent hypotheses that need to be verified, yet these annotations are accepted by databases and are included in the description line of the corresponding database entries. Obviously, referees cannot possibly review every single annotation provided by the authors, and this is where the classical peer-review system collapses, as indeed it does for most high-throughput, large-scale biological experiments. Alternative solutions must be sought.

Many open questions remain in our attempts to understand the science and art of genome re-annotation. To what degree can improved function predictions after genome re-annotation be assigned to growth in the public sequence databases, as opposed to the use of improved sequence-analysis algorithms, or simply to the use of different sequence-analysis algorithms (an algorithm such as BLAST is heuristic and will miss some similarities - use of a different heuristic algorithm may find other similarities), or to different human expertise? Can researchers in the genome-annotation field develop a gold-standard set of proteins to permit more objective evaluation of new automated analysis systems? How much variability is there in the annotations made by expert scientists? How much (if any) better are expert scientists than purely automated programs for genome annotation?

We have not discussed here various issues related to gene structure prediction, the differences between a variety of semi- and fully-automatic protocols for genome annotation, exchange formats and tools for the dissemination of information and the process of incorporating genome annotations in public databases. Undoubtedly, technological developments will assist towards our continually improving capacity for detecting functions encoded in genome sequences. These will, however, always have to be endorsed by the scientific community in order to have a real impact in the quality and scope of properly curated genome information resources of the twenty-first century.

## References

1. Ouzounis C, Kyrpides N, Sander C: **Novel protein families in archaean genomes.** *Nucleic Acids Res* 1995, **23:**565-570.
2. Editorial: **Debates over credit for the annotation of genomes.** *Nature* 2000, **405:**719.
3. Smith TF, Zhang X: **The challenges of genome sequence annotation or "the devil is in the details".** *Nat Biotechnol* 1997, **15:**1222-1223.
4. Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C: **Automated genome sequence analysis and annotation.** *Bioinformatics* 1999, **15:**391-412.
5. Iliopoulos I, Tsoka S, Andrade MA, Janssen P, Audit B, Tramontano A, Valencia A, Leroy C, Sander C, Ouzounis CA: **Genome sequences and great expectations.** *Genome Biol* 2000, **2:**interactions0001.1-0001.3.
6. **The European Bioinformatics Institute Computational Genomics Group** [http://www.ebi.ac.uk/research/cgg/services/]
7. Karp PD, Ouzounis C, Paley S: **HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*.** *Proc Int Conf Intell Syst Mol Biol* 1996, **4:** 116-124.
8. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*.** *Curr Biol* 1996, **6:**279-291.

9. Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C: **Challenging times for bioinformatics.** *Nature* 1995, **376:**647-648.
10. Robison K, Gilbert W, Church GM: **More *Haemophilus* and *Mycoplasma* genes.** *Science* 1996, **271:**1302-1303.
11. Ouzounis C, Casari G, Valencia A, Sander C: **Novelties from the complete genome of *Mycoplasma genitalium*.** *Mol Microbiol* 1996, **20:**898-900.
12. Huynen M, Snel B, Lathe WI, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210.
13. Andrade M, Casari G, de Daruvar A, Sander C, Schneider R, Tamames J, Valencia A, Ouzounis C: **Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function.** *Comput Appl Biosci* 1997, **13:**481-483.
14. Kyrpides NC, Olsen GJ, Klenk H-P, White O, Woese CR: ***Methanococcus jannaschi* genome: revisited.** *Microb Comp Genomics* 1996, **1:**329-338.
15. Kyrpides NC, Ouzounis CA: **Whole-genome sequence annotation: 'going wrong with confidence'.** *Mol Microbiol* 1999, **32:**886-887.
16. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomics: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25:**619-637.
17. Raghavan S, Ouzounis CA: **Novel coding regions in four complete archaeal genomes.** *Nucleic Acids Res* 1999, **27:**4405-4408.
18. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, Doerks T, Sanchez-Pulido L, Snel B, Suyama M, *et al.*: **Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames.** *Nucleic Acids Res* 2000, **28:**3278-3288.
19. Kyrpides NC, Ouzounis CA, Iliopoulos I, Vonstein V, Overbeek R: **Analysis of the *Thermotoga maritima* genome combining a variety of sequence similarity and genome context tools.** *Nucleic Acids Res* 2000, **28:**4573-4576.
20. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: **What's in a genome?** *Nature* 1992, **358:**287-287.
21. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: **Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III.** *Protein Sci* 1992, **1:**1677-1690.
22. Koonin EV, Bork P, Sander C: **Yeast chromosome III: new gene functions.** *EMBO J* 1994, **13:**493-503.
23. Andrade MA, Daruvar A, Casari G, Schneider R, Termier M, Sander C: **Characterization of new proteins found by analysis of short open reading frames from the full yeast genome.** *Yeast* 1997, **13:**1363-1374.
24. Ouzounis C, Bork P, Casari G, Sander C: **New protein functions in yeast chromosome VIII.** *Protein Sci* 1995, **4:**2424-2428.
25. Tsoka S, Promponas V, Ouzounis CA: **Reproducibility in genome sequence annotation: the *Plasmodium falciparum* chromosome 2 case.** *FEBS Lett* 1999, **451:**354-355.
26. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV: **Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs).** *Genome Biol* 2000, **1:**research0009.1-0009.19.
27. Pallen M, Wren B, Parkhill J: **'Going wrong with confidence': misleading sequence analyses of CiaB and clpX.** *Mol Microbiol* 1999, **34:**195.
28. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA: **Prediction of Enzyme Classification from protein sequence without the use of sequence similarity.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**92-99.
29. Shah I, Hunter L: **Predicting enzyme function from sequence: a systematic appraisal.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**276-283.
30. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol* 2000, **297:**233-249.
31. Bono H, Ogata H, Goto S, Kanehisa M: **Reconstruction of amino acid biosynthesis pathways from the complete genome sequence.** *Genome Res* 1998, **8:**203-210.
32. Tsoka S, Ouzounis CA: **Recent developments and future directions in computational genomics.** *FEBS Lett* 2000, **480:**42-48.
33. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15:**132-133.
34. Kyrpides NC, Ouzounis CA: **Errors in genome reviews.** *Science* 1998, **281:**1457-1457.
35. Venclovas C, Zemla A, Fidelis K, Moult J: **Some measures of comparative performance in the three CASPs.** *Proteins* 1999, **Suppl 3:**231-237.
36. Devos D, Valencia A: **Intrinsic errors in genome annotation.** *Trends Genet* 2001, **17:**429-431.
37. **The Institute of Genome Research** [http://www.tigr.org]
38. Lewis S, Ashburner M, Reese MG: **Annotating eukaryotic genomes.** *Curr Opin Struct Biol* 2000, **10:** 349-354.
39. Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW: **Functional and structural genomics using PEDANT.** *Bioinformatics* 2001, **17:**44-57.
40. **SWISS-PROT** [http://www.expasy.ch/sprot/sprot-top.html]