

Characterization and re-annotation of common genes found in 35 complete chloroplast genomes

Beatrice Kilel, (George Mason University), bkilel@gmu.edu

Abstract

The recent upsurge in complete chloroplast genomes calls for a renewed focus in ensuring that the information in the public repositories is properly annotated and cited. Annotation of new genomes as well as re-annotation of existing genomes is therefore a very important step in any sequence analysis. With well annotated sequences, evolutionary studies can be performed with a lot more certainty. Phylogenetic studies are based on the principle that if the sequences are similar, the genes to which they belong to must also be similar and tend to be functionally linked. This study focused on the re-annotation and characterization of genes found in 35 currently complete chloroplast genomes. Results obtained from re-annotation indicate drastic changes (over 30%) in the number of genes originally identified in most of the species which encoded proteins and functional genes. Frameshifts AT-rich “hot-spots” and point mutations in the sequence could explain this striking difference. *Adiantum capillus-veneris* sequence had 24% of photosystem II proteins before re-annotation and after re-annotation there were 42% proteins specific to this species and did not show any significant homology with BlastP searches.

The homologs of *Psilotum nudum* orf83 or orf119 could not be located in *Adiantum capillus-veneris* chloroplast genome. These may be spurious open reading frames or sequence divergence has lowered the similarity to a level that they are not detectable by Blast. The protein coding genes originally annotated as photosystem II have been changed and there are new hypothetical proteins. There are protein coding genes for ribulose 1, 5 - biphosphate carboxylase (Rubisco) and replacements of ORF 6 and photosystem II protein coding genes. Any ribosomal protein deletions may indicate that the function of these genes could be taken on by genes encoded in the nuclear genome.

Phylogenetic gene characterization studies indicate 3 distinct groups of non-green algae, green algae, and terrestrial plant specie based on parsimony distances. This aggregation indicates that during evolution, functional parts of DNA or protein sequences are under selective pressure, so they tend to evolve slower and are generally more highly conserved than non-functional sequences. Any local sequence conservation may indicate biological functionality. Knowing which genes is not found in a particular branch of the evolutionary tree and inferring when rearrangements occurred has enormous implications in plant regulatory network and general metabolism.

Introduction

Genome annotation is the process of adding information to the DNA sequence. The features added are either repeats, genes, promoters, terminators, splice sites or protein domains, which can be linked to other databases like Pfam/PubMed. There are several naming protocols and evidence codes in currently in place to allow for consistency during annotation. Artemis (Rutherford et al. 2000), Manatee (TIGR) software have been used successfully to store and view annotations using search data, paralogous families, and annotation suggestions generated from automated analysis. CDS (Snvder and Stormo, 1993) that include pseudogenes are thought of as a linear series of the sequence features: the initiation codon, coding sequence (exon), splice donor (5'), non-coding sequences (intron), splice acceptor (3'), coding sequence (exon) and termination codon. Similarity searches are performed on transcript mapping (ESTs and mRNAs), peptide mapping (nr protein databases and genomic mapping on other genomes). Important features to look for in the annotation process are transmembrane proteins, signal peptides, similarity with other proteins, low complexity regions, protein domains. It is important however to take the results obtained from annotation and protein sequences and subject them to molecular analysis and other comparative analysis techniques (Frazer et al. 2003; Kilel, 2003). Annotation features are more useful if protein function is attached and linked to the database to get more information through cross-references. There are several advantages for annotation such as cataloguing of model organisms as in gene ontology (GO) consortium (Ashburner et al. 2000) and Clusters of Orthologous Groups (COGs) (Tatusov et al. 2000). This structured choice of gene ontology terms will constrain use of nonsensical descriptions of gene products.

The re-annotation process is essential in any sequence analysis for the review of the coding sequences, updating and citing of current data, postulating functions, and making name changes (Bocs et al. 2002). To reduce the error procreation in the re-annotation process, several tools need to be used to compare the results obtained from different automatic annotation tools (Gaasterland et al., 2000; Burset and Guigo, 1996). It is important to have a manual review of data for concordance with transcript data, peptide similarity data as well as splice site usage (intron/exon boundaries).

Gene function prediction is an important process in the annotation of genomes in that it helps the researcher to decide the kind of experiments that are useful to a particular species or a particular gene (Burge and Karlin, 1997). Similarity with known peptides is an important method used in gene prediction (Pellegrini et al., 1999). Orthologs and paralogs, BlastP, and multiple sequence alignment (Delcher et al. 1999; Delcher et al. 2002) provide a basis for gene function prediction through homologous searches. Since most of the results are only predictions, the results should be taken as such and other supporting studies need to be performed such as evolutionary relationships in close juxtaposition with gene order and gene content (presence or absence of genes in a genome). Chloroplast genomes have been used successfully to study phylogeny because they are multifunctional (Stoebe and Kowallik, 1999; Wang, 2001) and they also have a small genome size (120-220 kb) (Hamby and Zimmer, 1991) and a few numbers of unique genes (120-150).

The study of genes vis-à-vis the location of the species on the phylogenetic tree is emerging as an important field of study in correlating how genomes genes have retained their structure due to minimal mutations, duplication, and any major rearrangement over the evolutionary distance (Gray, 1989; Eisen, 2000; Bansal and Meyer, 2002; Fitz-Gibbon, 1999). This will enhance and hasten a better understanding on the genomes that have not been fully sequenced but are closely related to the species that have been better studied. Identification of gene duplications can give insightful knowledge on the mechanism of gene duplication between the genomes. Locating the presence or loss of genes in the nucleus that have been transferred from the organelles can be best done through phylogenetic analysis (Sankoff and Nadeau, 2000). The significance of this study will help to explain the evolutionary changes in genomes in the prokaryotes and eukaryotes. This will help to provide a better understanding of genome rearrangements, location of homologous genes in the genomes that have not yet been sequenced, and thus investigate mechanisms of genome evolution.

Methods

GlimmerM automatic annotation software (Salzberg et al. 1999; Pertea and Salzberg, 2002) was used to re-annotate the 35 complete chloroplast genomes obtained from the GenBank (Benson et al. 1998).

This prediction tool is trained for *Arabidopsis thaliana*, *Oryza sativa* (rice) unlike other exon prediction tools that are trained for human and *Drosophila melanogaster*. It also has the strength of the splice sites and the score of the exons generated by an interpolated Markov model (IMM). Gene predictions from GlimmerM were compared to other annotation tools like Genotator (Harris, 1997). Results of GlimmerM exon prediction were submitted into Artemis annotation software (Rutherford et al. 2000), producing graphical genome annotation with mapped exons, CDS and proteins within the context of the genomic sequence and its six-frame translation. BlastP algorithm was used for homology searches and function prediction. Putative gene function assignments were performed manually.

Gene characterization was performed on 19 - gene data sets that are common to most of the complete chloroplast genomes obtained from the GenBank (Benson et al. 1998). Complete multiple sequence alignment was performed using ClustalX software (Thompson et al. 1997) and presented in the database as .GDE files. Parsimony analyses were performed through WINCLADA (Nixon, 1999a, b) and NONA algorithm (Goloboff, 1994). A heuristic search was carried out with 500 -1000 replicates, in order to obtain the most parsimonious solution with two starting trees per rep, multiple TBR + TBR (Tree Bisection and Reconnection) in order to obtain the optimal tree, and gaps treated as missing. Any uninformative characters were selected (using MOP command) and deleted. Jackknife and Bootstrap resampling methods were used to test the robustness of the various nodes on the tree topology. A consensus tree of 100-1000 replicas was utilized in this study with nodes occurring greater than 50% of the time being listed.

Results and Discussion

The results indicate a drastic change in the number of proteins after re-annotation in all the genomes as shown in Table 1. The reason for this major difference could be due frame-shifts or point mutations in the AT-rich “hot-spots” during the annotation process, which lead to wrong inferences in function prediction. There was a striking decrease (over 30%) in the proteins in pre-annotation and post-annotation sequences (Table 1). For concrete gene predictions, a manual review is needed for concordance with transcript data, peptide similarity data as well as splice site usage (intron/exon boundaries). Following re-annotation, there were four additional protein encoding genes involved in transport and Photosystem II (PSII) complex and with no change in the number of RNA genes in *Triticum aestivum*. It was also discovered that some protein-encoding genes originally annotated as photosystem II genes was re-assigned to another functional classes, while new photosystem II- related proteins emerged after re-annotation.

Five functional groups emerged from gene prediction results. These are mainly those used in photosynthesis, metabolism, transport, transcription/translation, and protein kinases or phosphatases. Before re-annotation the following functional groups existed in most of the genomes; NADH-plastoquinone oxidoreductase chain 1, maturase, photochlorophyllide reductase 1, ATP synthase CF1 alpha chain 6, ribosomal protein 16, cytochrome b6-f complex subunit 5, Acetyl coA carboxylase transferase 1, ATP binding protein 1, DNA-directed RNA polymerase alpha chain 3, photosystem I P700 chlorophyll A apoprotein A1 5, photosystem II 44 kDa apoprotein 15, and ORF 6. Following re-annotation all of the functional groups were maintained but photosystem II was missing. Ribulose 1,5-bisphosphate carboxylase large subunit, no significant homology, hypothetical protein, unknown protein, ycf protein, photosystem II, ORFQ-plastoquinone, elongation factor Tu, heat shock - B protein and chaperonin GroEL/GroES were introduced. The large subunit of chaperonin GroEL is present in all the non-green algae, which has disappeared in the higher terrestrial plants.

Branch support using the jackknife (and bootstrap - results not shown) program of Winclada and Nona gave results that were mostly above 75%, which is a good indication of the robustness of the tree topology that was generated (Figure 1). The jackknife test basically detects the differences in the protein domains (Apweiler et al. 2001) and show how the syntenic regions have been conserved over evolutionary distance. Parsimony jackknife scores are highest for *Pinus thunbergii* and *Pinus Koraiensis* (pinaceae) *Psilotum nudum* and *Adiantum capillus-veneris*, *Atropa belladonna* and *Nicotiana tabacum*, and the small grains *Triticum aestivum*, *Oryza sativa* and *Zea mays* (poaceae) with values at 100. *Oenothera elata* subsp. *hookeri*, *Arabidopsis thaliana*, and *Lotus japonicus* and *Spinacea oleracea* gave a collapsed clade with values at 66 at the base. Another collapsed clade was observed in *Odontella sinensis*, *Guillardia theta* and *Atropa belladonna*. Besides the clusters that did not receive greater than 60% support values were observed between *Astasia longa* and *Euglena gracilis*, *Nephrolepis olivacea* and *Chlorella vulgaris*, with values at

51 and 53 weak support values and a large clade at the base of the tree. This could be a result of a long conserved gene order and content and only recent changes noticed at the tips, indicative of a recent divergence between the two species (as shown in earlier studies De Las Rivas et al. 2002). Collapsed clades were obtained between *****. This is a common and intuitive technique, which can be used in phylogenetics since subtrees can be clustered with their common ancestor and thus collapsed and expanded. The results obtained need to be subject to different phylogenetic tools since deletions of uninformative characters of aligned sequences (Table 2) may affect the conclusions obtained from a phylogenetic analysis.

Table 1 Changes to protein coding genes following re-annotation of 35 complete chloroplast genomes

Genome	CDS (Before)	CDS (After)	Genes/Kb	Coding%
<i>Amborella trichopoda</i>	62	28	0.286	18.0
<i>Adiantum capillus-veneris</i>	89	36	0.239	10.7
<i>Anthoceros formosae</i>	88	47	0.291	8.2
<i>Arabidopsis thaliana</i>	87	18	0.116	11.3
<i>Astasia longa</i>	27	22	0.229	10.5
<i>Atropa belladonna</i>	92	40	0.255	20.5
<i>Calycanthus fertilis</i> var. <i>ferax</i>	88	33	0.215	15.8
<i>Chaetosphaeridium globosum</i>	99	52	0.396	26.0
<i>Chlamydomonas reinhardtii</i>	69	30	0.320	21.0
<i>Chlorella vulgaris</i>	70	62	0.405	21.4
<i>Cyanophora paradoxa</i>	145	60	0.442	26.1
<i>Cyanidioschyzon merolae</i>	100	30	0.214	14.5
<i>Cyanidium caldarium</i>	203	62	0.375	25.2
<i>Eimeria tenella</i>	28	19	0.076	4.98
<i>Epifagus virginiana</i>	31	12	0.171	22.9
<i>Euglena gracilis</i>	57	28	0.195	7.9
<i>Guillardia theta</i>	98	68	0.375	27.8
<i>Lotus japonicus</i>	80	48	0.318	21.5
<i>Marchantia polymorpha</i>	54	33	0.272	25.0
<i>Medicago truncatula</i>	Not in NCBI	22	0.177	20.5
<i>Mesostigma viride</i>	100	22	0.185	21.5
<i>Nephroselmis olivacea</i>	99	72	0.358	21.9
<i>Nicotiana tabacum</i>	87	37	0.237	18.8
<i>Odontella sinensis</i>	125	49	0.409	25.8
<i>Oenothera elata</i> subsp. <i>Hookeri</i>	80	36	0.219	18.4
<i>Oryza sativa</i>	59	31	0.230	8.8
<i>Physcomitrella patens</i> subsp. <i>Patens</i>	87	35	0.438	28.0
<i>Pinus koraiensis</i>	163	35	0.299	16.8
<i>Pinus thunbergii</i>	58	46	0.384	20.2
<i>Porphyra purpurea</i>	204	83	0.434	24.0
<i>Psilotum nudum</i>	83	36	0.259	18.7
<i>Spinacea oleracea</i>	79	49	0.325	19.8
<i>Toxoplasma gondii</i>	15	9	0.257	13.6
<i>Triticum aestivum</i>	85	33	0.245	10.6
<i>Zea mays</i>	73	18	0.128	8.3

Table 2 MOP uninformative characters of aligned sequences

Gene Name	Total # of characters	Characters deleted	# of characters used
atpA	2687	1825	862
atpB	3241	2494	747
psbA	2972	2533	439
psbC	12445	11797	648
psbD	10954	10511	443
psbE	928	800	128
psbF	1285	1209	76
rpl14	489	219	270
rpl16	1729	559	1170
rpl2	1706	785	921
rpl20	410	111	291
rpl36	230	157	73
rpoB	5036	2365	2671
rps11	608	299	309
rps14	413	177	236
rps2	1456	840	616
rps3	1237	628	609
rps4	760	266	494
rps8	1153	818	335

atpA-ATP synthase CF1 alpha chain, *atpB*-ATP synthase CF1 beta chain, *psbA*-Photosystem II Q(b) protein (D1), *psbC*- Photosystem II 44 kDa apoprotein (P6), *psbD*-Photosystem II D2 protein, *psbE*-Cytochrome b559 alpha chain, *psbF*-Cytochrome b559 beta chain, *rpl2*-50S ribosomal protein L2, *rpl14*-50S ribosomal protein L14, *rpl16*-50S ribosomal protein L16, *rpl20*-50S ribosomal protein L20, *rpl36*-50S ribosomal protein L36, *rpoB*-DNA-directed RNA polymerase beta chain, *rps2*-30S ribosomal protein S2, *rps3*-30S ribosomal protein S3, *rps4*-30S ribosomal protein S4, *rps8*-30S ribosomal protein S8, *rps11*- 30S ribosomal protein S11, *rps14*- 30S ribosomal protein S14

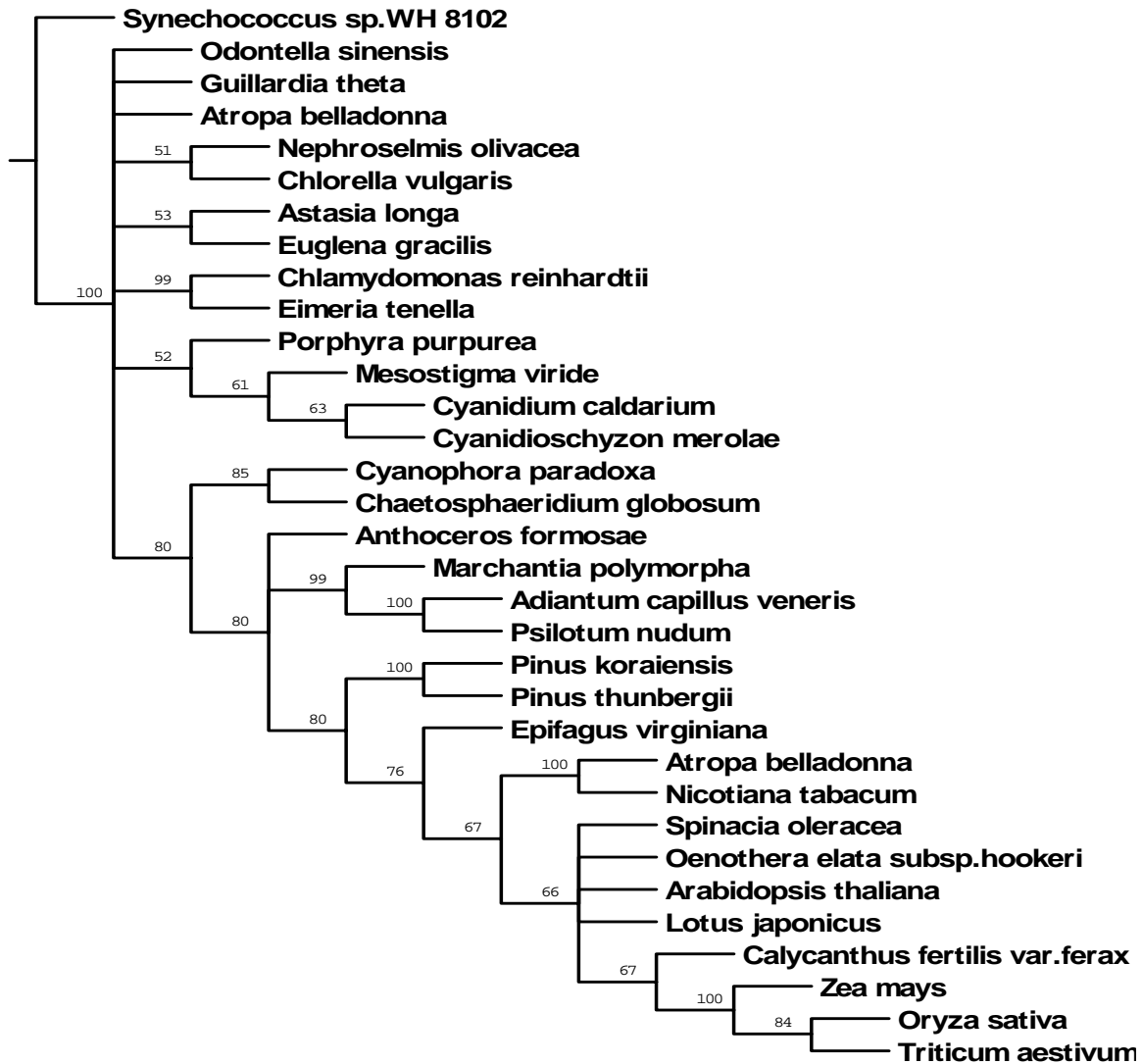


Figure 1 Consensus of most parsimonious trees with Jackknife support values placed above the tree branches. Generating a good tree is important in the inference of information across the genomes by comparative analysis.

Conclusions

Re-annotation of the current genomes is important since most of them are disparate and this may spawn errors in gene prediction. This is especially crucial for more conclusive comparative studies. Results in this study indicate new gene function assignments and change in function, which have been studied more using the current literature citations. There were new hypothetical proteins discovered, which would be subjected to further molecular techniques to add any extra data to the information already known. Automation of the annotation process is important if time and manpower required in the annotation process is to be curtailed. Phylogenetic studies are useful in explaining the nature of conserved genes which may result from lateral gene transfer (LGT) occurring between close and distant relatives (Nelson et al., 1999, Moret et al., 2001). With well annotated sequences, the knowledge obtained from one species can be extrapolated to close species. The new field of phylogenomics, which requires the generation of a proper phylogenetic tree in close association with the genes conserved will prove to be essential in the future genome evolutionary related studies.

References

- Apweiler, R., Attwood, T.K. et al. (2001), "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Res.*, 29, 37-40.
- Ashburner, M. and 20 other authors (2000), "Gene Ontology: tool for the unification of Biology". doi:10.1038/75556 25(1), 25 - 29.
- Bansal, A.K., and Meyer, T.E. (2002), "Evolutionary Analysis by Whole-Genome Comparisons," *J Bacteriol.*, 184, 2260-2272.
- Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., and Ouellette, B.F. (1998), "GenBank", *Nucleic Acids Research*, 26, 1-7.
- Bocs, S., Danchin, A., and Medigue, C. (2002), "Re-annotation of genome microbial CoDing-Sequences: finding new genes and inaccurately annotated genes," *BMC Bioinformatics*, 3, 1-5.
- Burge, C. and Karlin, S. (1997), "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol.*, 268, 78-94.
- Burset, M., and Guigo, R. (1996), "Evaluation of gene prediction programs," *Genomics* 34(3), 353-367.
- De Las Rivas, J., Lozano, J.J., and Ortiz, A.R. (2002), "Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns," *Genome Res* 12(4), 567-83.
- Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999), "Alignment of whole genomes," *Nucleic Acids Research*, 27(11), 2369- 2376.(MUMmer 1.0).
- Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002), "Alignment of whole Genomes," *Nucleic Acids Research*, 30(11), 2478-2483. (MUMmer 2).
- Eisen, J.A. (2000), "Assessing evolutionary relationships among microbes from whole-genome analysis," *Curr Opin Microbiol* 3, 475- 480.
- Fitz-Gibbon, S., and House, C.H. (1999), "Whole genome-based phylogenetic analysis of free-living microorganisms," *Nucleic Acids Res*, 27, 4218- 4222.
- Frazer K.A, Elnitski, L, Church, D.M, Dubchak, I., and Hardison R.C. (2003), "Cross species sequence comparisons: a review of methods and available resources," *Genome Res.*, 13, 1-12.
- Gaasterland, T., Sczyrba, A., Thomas, E., Aytekin-Kurban, G., Gordon, P., and Sensen, C.W. (2000), "MAGPIE/EGRET Annotation of the 2.9 Mb Drosophila melanogaster ADH Region," *Genome Research*, 10(4), 502-510.
- Goloboff, P. (1994), "Nona: A tree search program. Program and Documentation that is available from ftp.unt.edu.ar/pub/parsimony and www.cladistics.com.org."
- Gray, M. (1989), "The evolutionary origins of organelles," *TIG*, 5, 18.
- Hamby, R. K., and Zimmer, E.A. (1992), "Ribosomal RNA as a phylogenetic tool in plant systematics," In: *P. S. Soltis, D. E. Soltis, and J. J. Doyle (Eds) Molecular Systematics of Plants. Chapman and Hall, New York*, 50-91.
- Harris N. (1997), "Genotator: A Workbench for Sequence Annotation," *Genome Research* 7(7), 754 -762.
- Kilel, B. Analysis of chloroplast genomes databases and relationships using whole genome informatics tools. *PhD Diss.* George Mason University, School of Computational Sciences.
- Moret, B., Wyman, S., Warnow, T., and Wang, L. (2001), "New approaches for reconstructing phylogenies based on gene order," *Intelligent Systems for molecular biology*.
- Nelson K.E., Clayton RA., Gill S.R., Gwinn M.L., Dodson RJ, Haft DH, et al. (1999), "Evidence for lateral gene transfer between Archea and bacteria from genome sequence of *Thermotoga aritima*," *Nature*, 399, 323-329.
- Nixon, K.C. (1999a), "Winclada (beta) ver. 0.9.99m24," Published by the author, Ithaca, NY.
- Nixon, K.C. (1999b), "The parsimony ratchet, a new method for rapid parsimony analysis," *Cladistics* 15, 407- 414.
- Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999), "Assigning protein functions by comparative genome analysis: Protein phylogenic profiles," *Proc Natl Acad Sci., USA*, 96, 4285- 4288.
- Pertea, M., and Salzberg, S.L. (2002), "Using GlimmerM to find genes in eukaryotic Genomes," *Current Protocols in Bioinformatics*.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M-A, and Barrell, B. (2000), "Artemis: sequence visualisation and annotation." *Bioinformatics* 16 (10), 944-945.

- Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J., and Tettelin, H. (1999), "Interpolated Markov models for eukaryotic gene finding," *Genomics* 1, 24-31.
- Sankoff, D., and Nadeau, J.H. (2000), "Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families," Kluwer Academic publishers.
- Snyder, E.E., and Stormo, G.D. (1993), "Identification of Coding Regions in Genomic DNA Sequences: An Application of Dynamic Programming and Neural Networks," *Nucleic Acids Res.*, 21, 607 - 613.
- Stoebe, B., and Kowallik, K.L. (1999), "Gene-cluster analysis in chloroplast genomics," *TIG* 15, 344-347.
- Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000), "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Res.*, 28, 33-36.
- The Gene Ontology Consortium (2000), "Gene Ontology: tool for the unification of Biology," *Nature Genetics*, 25, 25-29.
- Thompson, J. D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997), "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Res.*, 25, 4876-82.
- Wang, L. (2001), "Exact-IEBP: A new technique for estimating evolutionary distances between whole Genomes," To appear, first workshop on algorithms in Bioinformatics (WABI'01), BRICS, University of Aarhus, Denmark, August 28-31.