Web Watch

SHOT: a web server for the construction of genome phylogenies

Jan O. Korbel*, Berend Snel*, Martijn A. Huynen and Peer Bork

With the increasing availability of genome sequences, new methods are being proposed that exploit information from complete genomes to classify species in a phylogeny. Here we present SHOT, a web server for the classification of genomes on the basis of shared gene content or the conservation of gene order that reflects the dominant, phylogenetic signal in these genomic properties. In general, the genome trees are consistent with classical gene-based phylogenies, although some interesting exceptions indicate massive horizontal gene transfer. SHOT is a useful tool for analysing the tree of life from a genomic point of view. It is available at http://www.Bork.EMBL-Heidelberg.de/SHOT.

The sequencing of genomes from cellular species has led to the development of methods that exploit the information from complete genomes to reconstruct phylogenies [1–3]. These methods use the number of shared orthologous genes or shared gene families between genomes as a similarity measure, rather than levels of sequence identity within a single gene

Box 1. Input parameters of SHOT

Gene-content phylogenies

Normalization to obtain the fraction of shared genes from the number of shared genes

(1) Division by the size of the smallest of the two genomes (theoretical maximum of shared orthologues).

(2) Division by the weighted average genome size (default selection). The weighted average is computed using $\sqrt{2}a \times b / \sqrt{a^2 + b^2}$ a fit to the number of orthologues shared between archaeal and bacterial genomes as function of the bacterial genome sizes (a and b are the sizes of both genomes; see Fig. 1 of Ref. [a]). This formula represents the data better than the genome size of the smaller genome, as the number of orthologues between Archaea and Bacteria also increases for large genomes – albeit slower.

Genome size definition

(1) Genome size is defined as the number of annotated protein coding open reading frames (ORFs).

(2) Genome size is the number of ORFs with at least one homologue in other genomes completed so far (default selection). Disregarding orphan ORFs eliminates considerable variation in gene prediction. It is therefore probably a better estimate of the maximum number of orthologues.

(3) Genome size is the number of ORFs with at least one orthologue in other completed genomes. This stringent option particularly affects genomes that experienced a high number of recent duplications. We recommend its use for investigating unexpected topologies, rather than as a standard option.

Distance measure

The evolutionary distance, d, is computed from the estimated similarity, s [b]

(1) $d = -\ln(s)$

(2) $d = 1 - s$

The default selection is function (1) because function (2) is less supported by models of evolution [b], hence providing a poorer estimate of evolutionary distances for weak similarities. However, function (2) can be applied for testing the robustness of clusters.

Clustering algorithm

(1) Neighbour-joining [c] (default selection).

(2) Fitch–Margoliash [d] (slower) can be applied instead.

Gene-order phylogenies

Genes considered for defining gene pairs

(1) ORFs annotated as genes are analysed for the presence of conserved gene pairs (default selection).

(2) Only genes shared between both genomes (ignoring genes without orthologues) are considered when defining gene pairs. Events that only affect the genomic gene content are ignored.

Normalization

Numbers of conserved gene pairs are normalized according to the genome size of the smaller genome (the maximum possible number of conserved gene pairs). Genome size can be defined as follows:

(1) number of ORFs annotated as genes;

(2) number of ORFs with at least one homologue in other complete genomes (default selection);

(3) number of ORFs with at least one orthologue in other complete genomes;

(4) In addition, the number of orthologues shared between two complete genomes can be used for normalization. We recommend applying this option when only shared orthologues are used for defining gene pairs.

Distance measure and clustering algorithm

Selectable parameters are identical to that of gene-content trees.

References

a Snel, B. *et al.* (1999) Genome phylogeny based on gene content. *Nat. Genet.* 21, 108–110

b Swofford, D.L. and Olsen, G.J. (1990) Phylogeny construction. In *Molecular Systematics* (Hillis, D.M. and Moritz, C., eds), pp. 411–501, Sinauer Associates Inc.

c Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for constructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425

d Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science* 155, 279–284

 0168-9525/02/$ – see front matter PII: S0168-9525(01)02597-5

family as has been done extensively; for instance, for small subunit ribosomal RNA [4,5]. Genome-based phylogenies are a welcome addition to gene-based phylogenies, because an unambiguous universal phylogeny based solely on comparisons within a single gene family seems unlikely [6]. Furthermore, complete genome trees are less affected by unrecognized horizontal gene transfer, unrecognized paralogy, highly variable rates of gene evolution, or misalignment than phylogenies based on single genes [1,2].

The construction of genome trees is not possible for everyone, as the comparison of complete genomes requires complex data processing and considerable CPU power. Thus, we have developed SHOT (for 'Shared Orthologue and gene-order Tree'), a construction tool that allows the generation of distance-based genome phylogenies on the web. Time-limiting genome comparisons are pre-computed and stored, allowing rapid online tree construction.

SHOT provides two independent strategies to construct trees:

(1) The gene content approach, in which the similarity between two genomes is the fraction of shared orthologous genes [1]. This method was refined by the incorporation of various options for calculation of the dissimilarity between genomes from the fraction of shared genes, including a new strategy for genome size normalization.

(2) SHOT also allows the generation of trees on the basis of gene-order conservation. Gene-order trees can be constructed only for prokaryotic genomes, as the order of genes in currently sequenced eukaryotes is too poorly conserved to contain a phylogenetic signal [7].

For both approaches, several parameter sets are available that can be selected depending on the type of question to be answered.

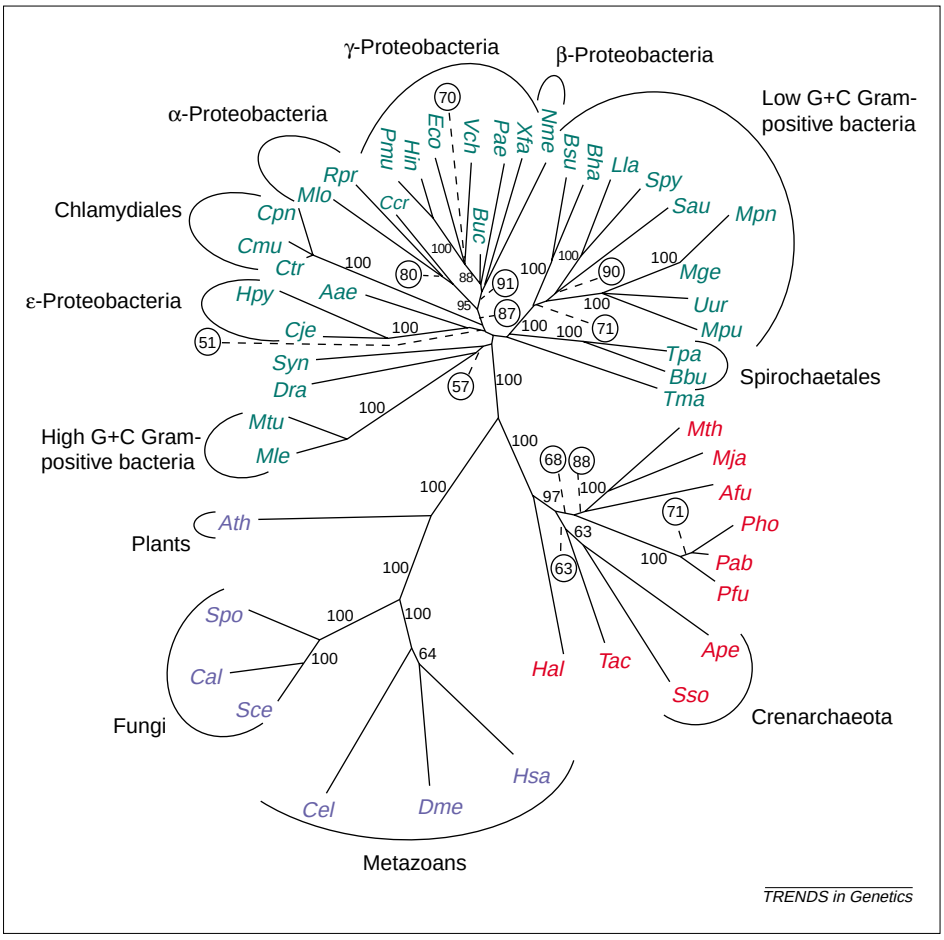

Fig. 1. SHOT gene content phylogeny based on 50 completed non-redundant genomes constructed applying the default parameters. Bootstrap values [1] of at least 50 (out of 100 replicates) are displayed to provide confidence estimates. Genomes considered (and the abbreviations used) encompass *Aeropyrum pernix* (*Ape*), *Aquifex aeolicus* (*Aae*), *Arabidopsis thaliana* (*Ath*), *Archaeoglobus fulgidus* (*Afu*), *Bacillus halodurans* (*Bha*), *Bacillus subtilis* (*Bsu*), *Borrelia burgdorferi* (*Bbu*), *Buchnera* sp. (*Buc*), *Caenorhabditis elegans* (*Cel*), *Campylobacter jejuni* (*Cje*), *Candida albicans* (*Cal*), *Caulobacter crescentus* (*Ccr*), *Chlamydia pneumoniae CWL029* (*Cpn*), *Chlamydia trachomatis* (*Ctr*), *Chlamydia muridarum* (*Cmu*), *Deinococcus radiodurans* (*Dra*), *Drosophila melanogaster* (*Dme*), *Escherichia coli K12* (*Eco*), *Halobacterium* sp. (*Hal*), *Haemophilus influenzae* (*Hin*), *Helicobacter pylori 26695* (*Hpy*), *Homo sapiens* (*Hsa*), *Lactococcus lactis* (*Lla*), *Methanobacterium thermoautotrophicum* (*Mth*), *Methanococcus jannaschii* (*Mja*), *Mesorhizobium loti* (*Mlo*), *Mycobacterium tuberculosis H37Rv* (*Mtu*), *Mycobacterium leprae* (*Mle*), *Mycoplasma genitalium* (*Mge*), *Mycoplasma pneumoniae* (*Mpn*), *Mycoplasma pulmonis* (*Mpu*), *Neisseria meningitidis Z2491* (*Nme*), *Pasteurella multocida* (*Pmu*), *Pseudomonas aeruginosa* (*Pae*), *Pyrococcus abyssi* (*Pab*), *Pyrococcus horikoshii* (*Pho*), *Pyrococcus furiosus* (*Pfu*), *Rickettsia prowazekii* (*Rpr*), *Saccharomyces cerevisiae* (*Sce*), *Schizosaccharomyces pombe* (*Spo*), *Staphylococcus aureus Mu50* (*Sau*), *Streptococcus pyogenes* (*Spy*), *Sulfolobus solfataricus* (*Sso*), *Synechocystis* sp. (*Syn*), *Thermoplasma acidophilum* (*Tac*), *Treponema pallidum* (*Tpa*), *Thermotoga maritima* (*Tma*), *Ureaplasma urealyticum* (*Uur*), *Vibrio cholerae* (*Vch*) and *Xylella fastidiosa* (*Xfa*).

Methodology, input and output

We use an operational definition of orthology to predict genes shared between genomes (for details, see Ref. [1]), namely considering non-overlapping bi-directional best hits in Smith–Waterman [8] protein sequence comparisons (E-value $\leq 10^{-2}$). For gene content phylogenies, the similarity between two species is defined as the ratio of the number of shared orthologues and a normalization value that reflects varying genome sizes. The normalization value is dominated by the number of genes in the smaller of the two compared genomes, because that is the number that determines the maximum number of genes two genomes can share. Independent, large-scale loss of genes, as is often observed in parasites, does therefore not lead to a clustering of such small genomes into one branch of the tree, because these small genomes still share more genes with their large, closest relatives (see results presented here and in Ref. [1] for examples), than with the other small genomes. Note that such co-clustering of small, distantly related genomes is indeed apparent in gene-content-based genome trees that do not normalize genome sizes in the manner implemented in SHOT and that also include the absence of genes to calculate genome similarity [2,3].

For gene-order phylogenies, similarities are derived from the number of orthologous gene pairs conserved. We define a 'conserved gene pair' as orthologous genes that in two genomes form an adjacent pair of genes with the same conserved relative directions of transcription. SHOT uses tools from the PHYLIP package [9] to construct phylogenetic trees.

As input of SHOT, a set of species is selected. The default output is an image

*These authors contributed equally.

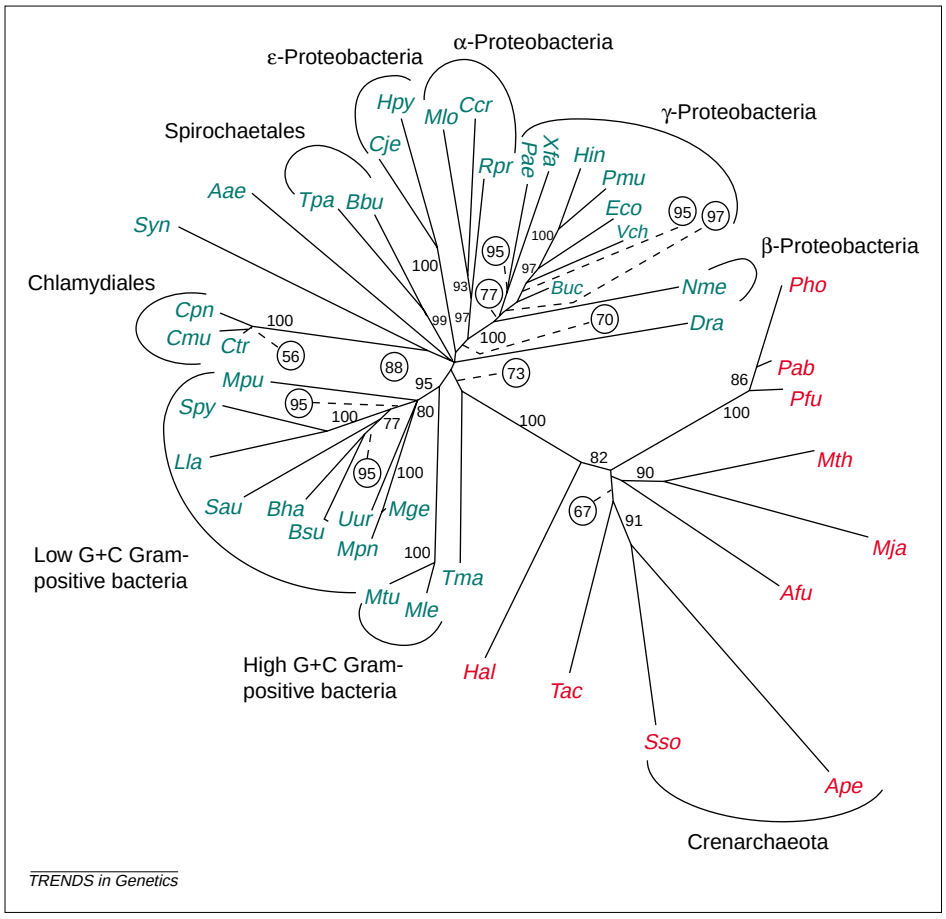

Fig. 2 SHOT gene-order phylogeny of all prokaryotic species listed in Fig. 1, using default parameters.

of an unrooted tree with the option to download the tree as a postscript file or in Newick format that is compatible with various phylogeny software packages. Among several adjustable parameters (see Box 1), the calculation of bootstrap values can be selected.

Genomic data were extracted from GenBank (ftp://ncbi.nlm.nih.gov), except for the species *Caenorhabditis elegans* (downloaded from ftp://ftp.sanger.ac.uk/pub/databases/wormpep), *Candida albicans* (ftp://cycle.stanford.edu/pub/projects/candida), *Arabidopsis thaliana* (ftp://ftp.arabidopsis.org/home/tair), *Schizosaccharomyces pombe* (http://www.sanger.ac.uk/Projects/S_pombe), and *Homo sapiens* (http://www.ensembl.org).

Comparison of SHOT trees with a small subunit rRNA tree

We discuss here some features of SHOT trees that have been constructed using all currently sequenced genomes of cellular species. In Figs 1–3, we present genome trees constructed using the two methods available in SHOT, along with a small subunit ribosomal RNA (SSU rRNA) tree generated using the RDP website [5]. Both the gene-content tree and the gene-order tree show a remarkable similarity with the SSU rRNA tree. Whereas a phylogenetic signal in gene content has been demonstrated previously [1–3], the results indicate that the conservation of gene order also reflects the evolutionary distances of the respective species. Both types of genome trees reveal clustering of several known clades of the tree of life with high bootstrap values – such as the metazoans and fungi, chlamydiae, spirochetes, low G+C Gram-positive bacteria, high G+C Gram-positives, and the α- and ε-proteobacteria. Of the trees presented here, only the gene-order tree separates the β- and γ-proteobacteria and reveals a monophyly of Gram-positive bacteria. Whether Gram-positive bacteria form a single monophyletic clade is still a matter of discussion [10].

SHOT should provide a helpful tool to shed new light on disputed points of the universal species phylogeny. For instance, the gene-content tree reveals *Homo sapiens*, and not *C. elegans*, as the closest sequenced metazoan relative of *Drosophila melanogaster*. This topology resembles the traditional animal phylogeny based on morphology and embryology, as well as newer phylogenies based on combined protein data [10,11], but not phylogenies based on SSU rRNA sequence identity, which reveal a clustering of *D. melanogaster* with *C. elegans* (see Ref. [12] and references therein).

The branching observed for the methanogenic Archaea, the pyrococci, and *Archaeoglobus fulgidus* differs significantly from a topology derived from rRNA, as previously discussed in detail [1]. The topology revealed earlier on a smaller set of genomes proved robust against the addition of new taxa, and is moreover supported by gene-order trees.

Impact of horizontal gene transfer

Sometimes the phylogenetic signal is obscured by horizontal gene transfer (HGT), which, for instance, causes problems in the rooting of the archaeal branch. Genome trees reveal *Halobacterium* sp. at the root of the Archaea, and clustering of the euryarchaeon *Thermoplasma acidophilum* with the two crenarchaeota *Aerophyrum pernix* and *Sulfolobus solfataricus*. We argue that this is the result of substantial HGT that occurred between *T. acidophilum* and *S. solfataricus* [13] as well as between *Halobacterium* and the Bacteria [14]. This assumption is supported by the finding that *Halobacterium* disappears from the root, when a gene-content tree without the Bacteria but with Archaea and Eukaryotes is constructed (not shown). Moreover, euryarchaeaota and crenarchaeota are monophyletic and correctly rooted, when a gene-content tree of all sequenced cellular genomes excluding *Halobacterium* and *T. acidophilum* is constructed (not shown).

In our opinion, such findings do not decrease the relevance of genome trees generated by SHOT. A main feature of genome phylogenies is that, rather than disclosing the history of single genes, they reflect the evolutionary history of complete genomes. Large numbers of horizontally transferred genes considerably affect the organisms' evolution and phenotype. Similarly, lifestyles might influence the gene content. Constructing genome-based phylogenies along with phylogenies produced by traditional tree reconstruction techniques is therefore relevant, as it

could be very helpful to visualize such peculiarities of genome evolution.

When should options of SHOT be applied?

Gene order evolves faster than gene content [15]. Hence, gene order phylogenies perform particularly well for short evolutionary distances. For instance, in contrast to gene-content trees, gene-order trees reveal *Staphylococcus aureus* at its consensus position within the low G+C Gram-positives as a sister species of *Bacillus subtilis*. For larger evolutionary distances, we recommend gene-content phylogenies. Generally, we suggest starting with the default parameters. However, as there is no accepted relevant model for genome evolution, alternative parameter selections can also result in meaningful phylogenies (see Box 1 for parameter effects). Bootstrapping and parameter changes can be applied to study the robustness of clusters, or the phylogeny of species for which signals provided by genomic gene content and gene order are other than phylogenetic.

For instance, the gene-order tree reveals elongated branch lengths for species such as *Synechocystis* sp. or *A. pernix*. *Synechocystis* has an extensively shuffled genome [15], whereas the genome of *A. pernix* appears to include a number of open reading frames that are incorrectly annotated as genes [16]. The branch length of the latter species significantly decreases, if genes not shared between two species are ignored when defining gene pairs (option 'shared orthologues' selected in the field 'Genes considered for defining gene pairs').

The results obtained for the thermophilic bacteria *Thermotoga maritima* and *Aquifex aeolicus*, both placed at the root of the Bacteria in SSU rRNA trees, provide examples of how to evaluate phylogenetic information obtained from changing parameters in SHOT. Using the default parameters, *A. aeolicus* clusters with the ε-proteobacteria in gene content trees. When the input parameters of gene content and gene order tree construction methods are varied, *A. aeolicus* clusters with the proteobacteria or appears close to the root of the Bacteria. *T. maritima* appears at the root of the Bacteria for many parameter selections in gene-order trees, but tends to cluster rather with the low G+C Gram-positives in gene-content trees. The placement of thermophiles at the bacterial root in single-gene trees might be an artefact owing to varying evolutionary rates in thermophiles compared with mesophiles [16,17], whereas in genome trees this is caused by massive HGT from Archaea to thermophilic Bacteria [18]. Thus, the recurring clustering of *A. aeolicus* with the ε-proteobacteria and *T. maritima* with the low G+C Gram-positives might reflect their true phylogeny.

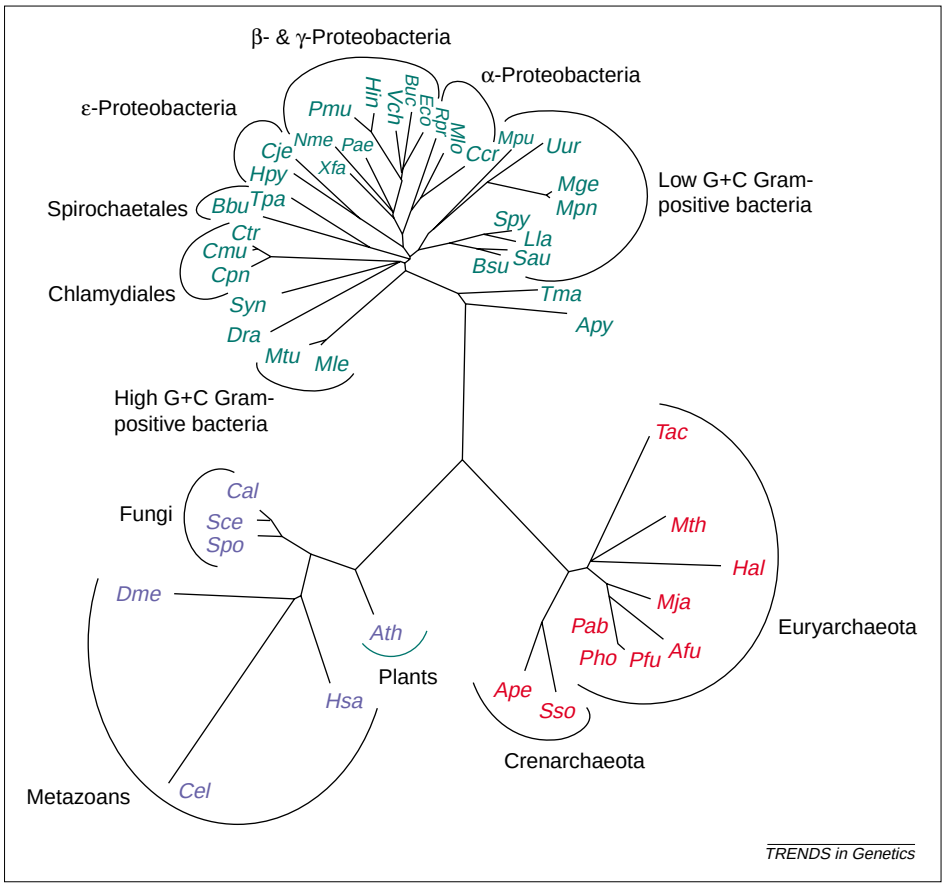

Fig. 3. Phylogeny of the species listed in Fig. 1, made on the basis of small subunit rRNA. A 16S rRNA tree of prokaryotes and an 18S rRNA tree of eukaryotes were constructed using the RDP website [5]. The eukaryotic subtree was added to the 16S rRNA tree at its consensus position, with *Arabidopsis thaliana* at the root. Note that the length of the branch leading to the eukaryotes is thus not necessarily correct. Because the 16S rRNA is not available for *Aquifex aeolicus* in RDP, we use the close relative *Aquifex pyrophilus* (*Apy*).

Conclusion

SHOT is a web server for the reconstruction of genome trees that calculates evolutionary distance from gene acquisition and loss, or from genome rearrangement, depending on which method is selected. Several groups have constructed phylogenetic trees from conserved gene orders of animal mitochondria (see Ref. [6] and references therein) and for particular clusters of bacterial genes [19,20]. However, as far as we know, we are the first to exploit gene-order conservation of whole genomes to construct trees of prokaryotes.

SHOT is updated constantly to include new genomes. The addition of more genomes should improve the robustness of results from SHOT and help to resolve disputed issues, in particular the clustering of species that still lack a sequenced close relative. We expect that instead of the complete set of available taxa, future studies will rather focus on selected subsets of species, allowing the study of phylogenies at different levels of resolution. Finally, SHOT should be useful not only for resolving conflicts on the basis of single-gene phylogenies, but also, by comparing genome-based phylogenies with single-gene phylogenies, for acquiring an overview of the evolution of basic genomic features, namely gene content and gene order.

Acknowledgements

This work was supported by BMBF. J.O.K. and P.B. also carry out research at Max Delbruck Centre for Molecular Medicine, Berlin-Buch, Germany. We thank Warren Lathe III and Steffen Schmidt for helpful comments on the manuscript.

Jan O. Korbel*
Berend Snel
Peer Bork
EMBL, Meyerhofstrasse 1,
69117 Heidelberg, Germany.
*e-mail: korbel@embl-heidelberg.de

Martijn A. Huynen
Nijmegen Center of Molecular Life Sciences, p/a CMBI, Toernooiveld 1,
6525 ED Nijmegen, Netherlands.

References

1 Snel, B. *et al.* (1999) Genome phylogeny based on gene content. *Nat. Genet.* 21, 108–110
2 Fitz-Gibbon, S.T and House, C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* 27, 4218–4222
3 Tekaia, F. *et al.* (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.* 9, 550–557
4 Olsen, G.J. *et al.* (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* 176, 1–6
5 Maidak, B.L. *et al.* (1997) The RDP (Ribosomal Database Project). *Nucleic Acids Res.* 25, 109–111
6 Boore, J.L. and Brown, W.M. (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.* 8, 668–674
7 Huynen, M.A. *et al.* (2001) Inversions and the dynamics of eukaryotic gene order. *Trends Genet.* 17, 304–306
8 Smith, T. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197
9 Felsenstein, J. (1989) PHYLIP-phylogeny inference package (Version 3.2). *Cladistics* 5, 164–166
10 Brown, J.R. *et al.* (2001) Universal trees based on large combined protein sequence data sets. *Nat. Genet.* 28, 281–285
11 Baldauf, S.L. *et al.* (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290, 972–977
12 Graham, A. (2000) Animal phylogeny: root and branch surgery. *Curr. Biol.* 10, R36–R38
13 Ruepp, A. *et al.* (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 407, 508–513
14 Ng, W.V. *et al.* (2000) Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12176–12181
15 Huynen, M.A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.* 54, 345–379
16 Cambillau, C. and Claverie, J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.* 275, 32383–32386
17 Forterre, P. (1998) Were our ancestors actually hyperthermophiles? Viewpoint of a devil's advocate. In *Thermophiles: The Keys to Molecular Evolution and the Origin of Life?* (Wiegel, J. and Adams, M.W.W., eds), pp. 137–146, Taylor & Francis Inc.
18 Nelson, K.E. *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329
19 Tamames, J. *et al.* (2001) Bringing gene order into bacterial shape. *Trends Genet.* 17, 124–126
20 Tamames, J. *et al.* (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol.* 2, 0020.1–0020.11 (www.genomebiology.com/)

Book Review

Eugenics, a good idea?

The Unfit: A History of a Bad Idea
by Elof Axel Carlson
Cold Spring Harbor Laboratory Press, 2001.
£30.00 hbk (451 pages) ISBN 0 8769 587 0

'A great deal of the literature of eugenics is based on a myth. The myth…pits the forces of evil and power…against the forces of innocence and vulnerability… The myth is wrong and dangerous,' says Carlson. He has tried to write a book to contradict the myth. The book is excellent. However, in my opinion, it just confirms this myth.

Positive eugenics, the science of encouraging the genetically superior to increase their reproduction, has been widely discussed but hardly ever practised. By contrast negative eugenics, the science defining the 'unfit' who have to be excluded from procreating became real. It has a long and complex history. So far, most of the history has been written by historians of science. However this history of negative eugenics is written by a student of the great geneticist Hermann Muller, and a geneticist himself.

He begins his story with the genocide performed on the Amalakites, which is recorded in the Bible. He does not mention the eugenic utopia proposed by Plato, but goes from the Bible straight to Europe of the 17th century. What a wonderful mosaic of ideas! Carlson does not rely on secondary sources; he has read the original documents and quotes them extensively. He presents a most lively panorama of the ideas historically used to justify negative eugenics. I list just some of them: masturbation as the cause of mental illness and the mentally ill should not procreate. Malthus's *Essay on the Principles of Population* [1], where he blames the poor for their misfortunes. Spencer's *Social Statics* [2], the founding document that led to what later became social darwinism. Gobineau's book on *The Inequality of Human Races* [3]. Morel's book on *Degeneracy* [4] and how Zola made use of this idea for his novels (e.g. Ref. [5]).

From there Carlson proceeds to Mendel and Darwin. Then comes Galton, who gave eugenics its name, and showed that his cousin Darwin's idea about inheritance was wrong. He details how sterilization of the unfit was proposed for the first time in the US, and how the first attempt to pass such a law failed, although the law was eventually passed in Indiana (1907). He gives accounts of the investigations of the American populations of the 'unfit', the Jukes, the tribe of Ishmael and the Kallikaks. Sometimes the details are little bit too detailed. For example, we read that Prescott Hall, a lawyer who helped Davenport with his attempt to limit the immigration of the unfit, listened to Wagner operas to treat his chronic insomnia (p. 257).

Most revealing is the story of the growth and decay of eugenics in the various countries. By 1940 in the USA, 35 878 persons had been sterilized. The immigration laws had been rewritten in favour of the Nordic. The Supreme Court had made the Buck vs. Bell decision, legalizing the sterilization of persons regarded as being feebleminded against their will. All this is described in great detail with illuminating citations. Curiously there is one aspect missing. Carlson does not mention the US laws that outlawed marriage and sexual intercourse between blacks and whites. In 1923, 28 states had such laws, which, in 1935, became the models for the German Nuremberg laws outlawing marriage and intercourse between Jews and non-Jewish Germans.

The consequences of negative eugenics were bad in the USA and the Scandinavian countries. But they were a disaster in Germany. There, the geneticists collaborated with the Nazis and legitimized their antisemitism. This development is condensed into one brief chapter in Carlson's book. Here, only secondary sources are cited. Excellent books are not mentioned; for example, Friedlander's *The Origins of Nazi Genocide: from Euthanasia to the Final Solution* [6]. Also, minor details are inaccurate: the Hitler Putsch happened in 1923 not 1924;