COMPARATIVE GENOME ANALYSIS: EXPLOITING THE CONTEXT OF GENES TO INFER EVOLUTION AND PREDICT FUNCTION

Peter Bork
Berend Snel
Gerrit Lehmann
Mikita Suyama
Thomas Dandekar
Warren Lathe III
Martijn Huynen

Different gene context methods are reviewed and summarized and an example of implementing several of these strategies into a publicly accessible web server is given.

1. Introduction

Although a lot of theoretical groundwork has been done on the evolution of gene order (Sankoff and Blanchette, 1999; Sankoff, 1993), genomes of free living organisms have only recently become available in sufficient quantity to allow multiple comparisons and infer evolutionary constraints. For example, it became clear that despite an enormous amount of gene shuffling in each phylogenetic branch there is both local and nonlocal conservation that appears to reflect a number of functional constraints on the gene products, the proteins. In the last three years, a number of approaches have been developed (Marcotte et al., 1999a,b; Enright et al., 1999; Overbeek et al., 1999; Pellegrini et al., 1999; Dandekar et al., 1998; Huynen and Bork, 1998) to exploit such gene context conservation to infer functional association of the respective proteins. Individual methods have been reviewed extensively (Huynen et al., 2000; Marcotte, 2000; Teichmann and Mitchison, 2000; Huynen and Snel, 2000; Doolittle, 1999; Sali, 1999). We give here an overview and summarize

D. Sankoff and J.H. Nadeau (eds.), Comparative Genomics, 281–294.

different gene context methods, give an example of implementing several of those strategies into a publicly accessible web server and also discuss evolutionary considerations associated with the conservation of gene context.

2. Gene context and knowledge context

Prediction of gene function currently means transfer of existing knowledge from a knowledge base, usually a sequence database with its annotations. This is bearing the danger that errors in our knowledge base are propagated or incorrect inferences are being made (e.g., Bork and Koonin, 1998). Currently, the accuracy of those methods is rarely better than 70% (Bork, 2000). Advantages of utilizing gene context information are that it is directly contained in the data and that the signal for recognition is increasing with the increasing amount of data (provided they are of high quality) i.e. power of comparative analysis increases in time. Gene context can be locally on the chromosome (e.g., conservation of gene neighborhood or gene fusion), can be scattered throughout the genome (e.g., shared or similar regulatory elements) or can only become visible on the background of multiple genomes (e.g. co-occurrence of genes in the same subset of complete genomes). All these methods can only be used for function prediction if one combines them with existing knowledge, the knowledge context. This is basically our computerized knowledge on biological processes accessible in databases such as for metabolic networks (e.g., KEGG, Kanehisa and Goto, 2000 WIT, Overbeek et al., 2000), for protein interaction (e.g., DIP, Xenarios et al., 2000), for cellular localization (Nakai, 2000; Eisenhaber and Bork, 1998), for expression experiments e.g., Aach et al., 2000; Scherf et al., 2000, for known post-translational modifications and so on. Furthermore, methods are now being developed to directly access primary sources for biological knowledge, the scientific literature e.g., Andrade and Valencia, 1998; Rebhan et al., 1998. In order to exploit the signal inherent to DNA for biological predictions, the knowledge base has to be accessed and again, at this stage, inference of functional features might introduce some noise into the prediction.

In the following, we introduce some of the concepts that are relevant for the usage of gene context information.

2.1. Gene order

Gene order has been described since some time to be not conserved in prokaryotic evolution (Kolsto, 1997) although some local conservation is retained (Mushegian and Koonin, 1996; Tamames et al., 1997; Watanabe et al., 1997). Gene order conservation is decreasing with the phylogenetic distance of the species compared (Huynen and Bork, 1998). **A** more quantitive comparison of nine genomes revealed that indeed the conservation of local gene neighborhood in three distant species allows the inference of direct or indirect (e.g., part of the same complex) physical interaction (Dandekar et al., 1998). If the criterium for local neighborhood is a bit relaxed, e.g. only the conservation of the presence within equivalent operons (termed "runs" as our knowledge on shared regulatory elements is still too fragmentary to infer the presence of operons) is required, the signal is still very strong (Overbeek et al., 1999, 2000). An example for deriving a hypothesis using context information is illustrated in Figure **1.** To quantify conservation of local gene neighborhood, one needs to take into account the phylogenetic distance of species, their genome sizes, the distribution of the neighboring genes in all the genomes considered, and effects such as horizontal gene transfer. The latter obviously creates noise as neighboring genes have not had enough time to become shuffled despite their presence in divergent species. In order to estimate the degree of genome distance that is required to assume almost complete genome shuffling, the average similarity of orthologous genes (i.e., genes that reveal speciation events; Fitch, 1970) or of ribosomal RNAs in the genomes compared can be taken as a measurement of phylogenetic distance (Huynen and Snel, 2000; Doolittle, 1999; Huynen and Bork, 1998). It can be compared to the degree of neighborhood conservation (Figure 2). As operons require the same transcription direction, the 5' to 3' arrangements should be the strongest conserved, a weak signal has also been observed for 5' to 5' arrangements, i.e., divergent promoters (Huynen and Snel, 2000) (Figure 3), while 3' to 3' arrangements appear the least conserved. At an average distance of 87% small subunit ribosomal RNA sequence identity or an amino acid sequence identity of 70% between all identifiable orthologous gene products of two genomes, conservation of gene order can already be seen as preserved due to functional constraints on the gene products. Note that this rule might only apply to prokaryotic genomes; poxvirus genomes seem to follow different rules (Figure 4).

Recently, a number of closely related genomes have become accessible and it is now also possible to gain insight into the mechanisms that lead to the genome shuffling, namely global and local inversions, gene and gene cluster duplications and loss as well as recombination events (Figure **4).** It becomes clear that the quantification of those events becomes difficult as different constraints apply to different lineages (e.g., missing inversions in the mycoplasma lineage might be due to the absence of palindromic sequences and plasmid integration sites which in turn might be due to the absence of restriction enzymes in these organisms (Gelfand and Koonin, 1997)).

2.2. Gene fusion

Another type of context is based on the assumption that genes that are fused do functionally associate or even physically interact. As orthologous gene products are likely to perform the same function in other organisms, the single occurrence of gene fusion in one organism is enough to predict interaction of the gene products (Enright et al., 1999). Indeed, the method seems very accurate with only very

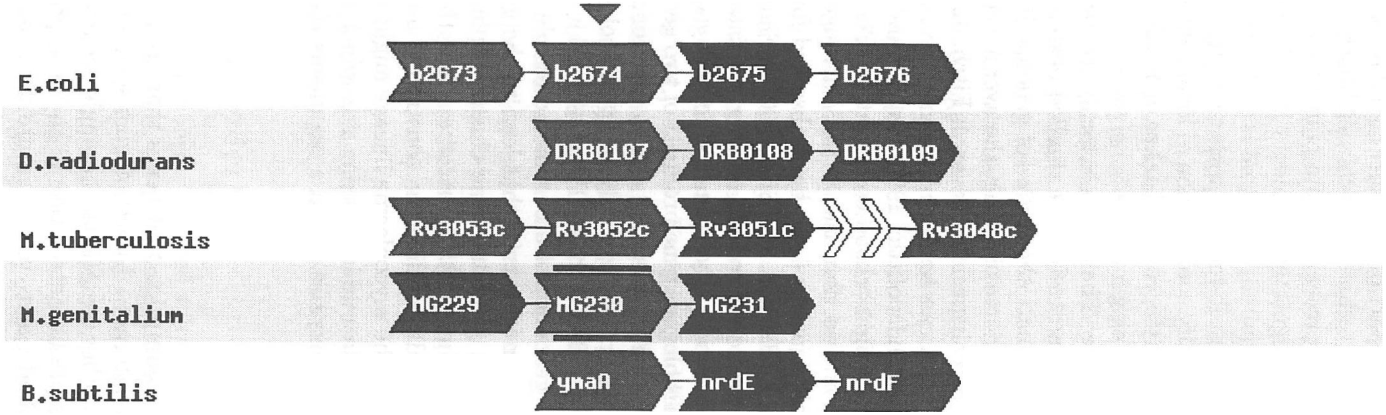

Figure 1. This figure shows the gene clusters in which nrdI (the red gene symbols) and nrdE (the black gene symbols) are present together. Now, with the background of completely sequenced genomes, one can construct predictions founded and based on this pattern. The genomic association of nrdI with the genes in the nucleotide reductase operon nrdE and nrdF (the brown gene symbols), has been observed before, and it has been shown that nrdI has a stimulatory effect on ribonucleotide reduction (Jordan 1997). Based on this pattern, we predict a functional association between the two gene products. The gene-order is conserved for all the published nrdI genes, including e.g. bacteriophage SPBc2 (although this is not shown in the figure). We therefore postulate a physical interaction between NrdI and NrdE.

Genes from the same orthologous family have the same color. The truncated small white gene-like symbols are genes that are be located between the genes retrieved via the conserved gene clusters, but that are themselves not conserved in that position. The gene symbols with two colors are assigned to different gene families, because they are the result of fusions. An interruption symbol means that the two displayed stretches of the genome are not in the same gene cluster. The lines between the genes symbolize the stretches of DNA in between the genes.

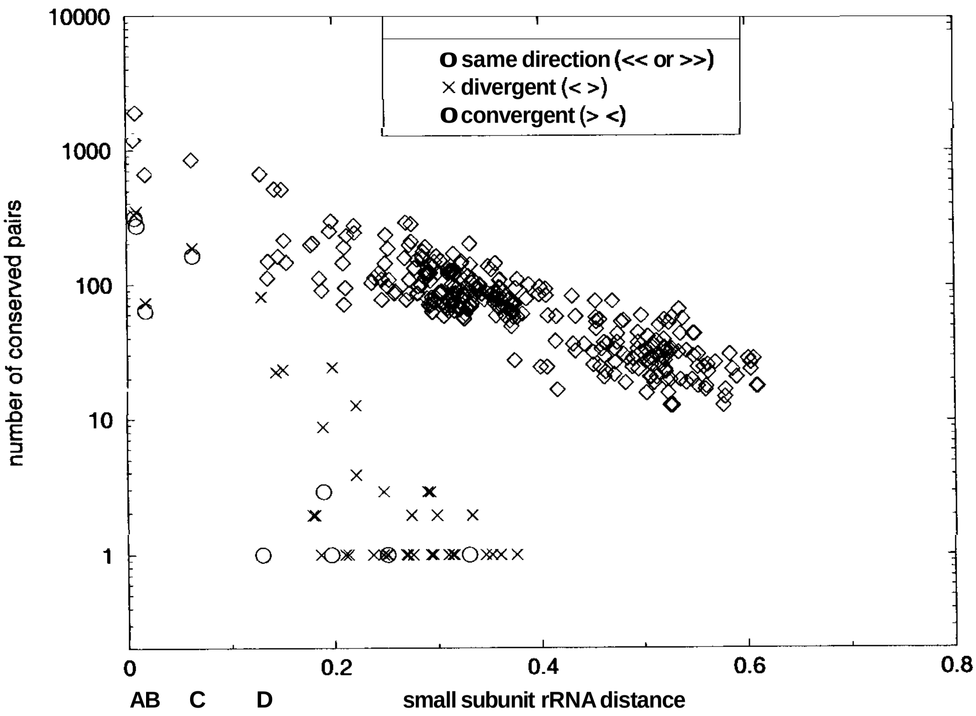

Figure 2. Adjacent genes with conserved relative transcription orientation versus evolutionary distance. Shown are gene pairs with a single direction of transcription, gene pairs with a convergent direction of transcription, and gene pairs with a divergent direction of transcription. The X-axis represents the small subunit rRNA distance between the species. The number of conserved adjacent pairs declines nonlinearly with phylogenetic distance and was therefore depicted on a logarithmic scale. Note that the vast majority of conserved adjacent pairs with a conserved relative transcription direction are transcribed in the same direction. Divergently transcribed pairs are better conserved than convergently transcribed ones, hinting at the conservation of divergent promotors. The distances A, B, C, D correspond to the genome pairs depicted in Figure 3 A, B, C, D respectively.

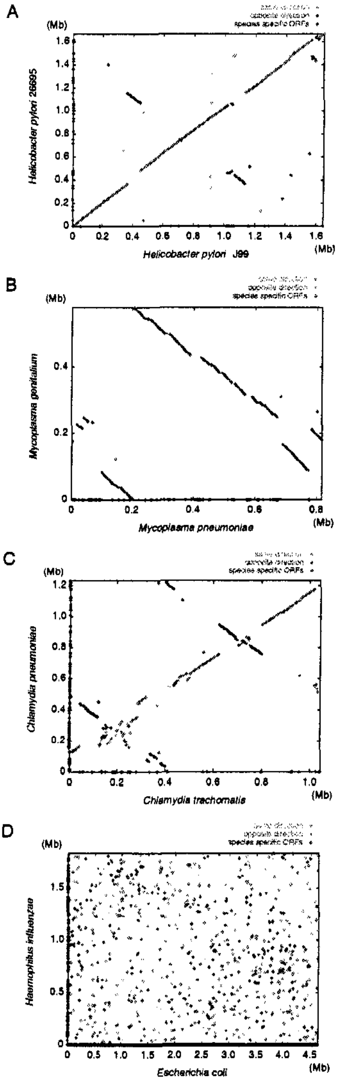

Figure 3. Dot plots of orthologous genes between closely related bacterial genomes: A, *Helicobacter pylori* J99 and *Helicobacter pylori* 26695; B, *Mycoplasma pneumoniae* and *Mycoplasma genitallium;* C, *Chlamydia trachomatis* and *Chlamydia pneumoniae;* D, *Escherihia coli* and *Haemophilus influenzae.* Orthology is defined as "bidirectional best, significant ($E < 0.01$) hit" based on Smith-Waterman (Smith and Waterman, 1981) comparisons of the genomes with one another and including the possibility of gene fusion/fission (Huynen and Bork, 1998). Directional similarity is indicated by colors: green, pairs of genes on the same direction; red, those on the opposite direction. The ORFs without significant similarity to the other compared genome even in local DNA sequence level are defined as the species specific ORF and indicated by blue dots on each axes.

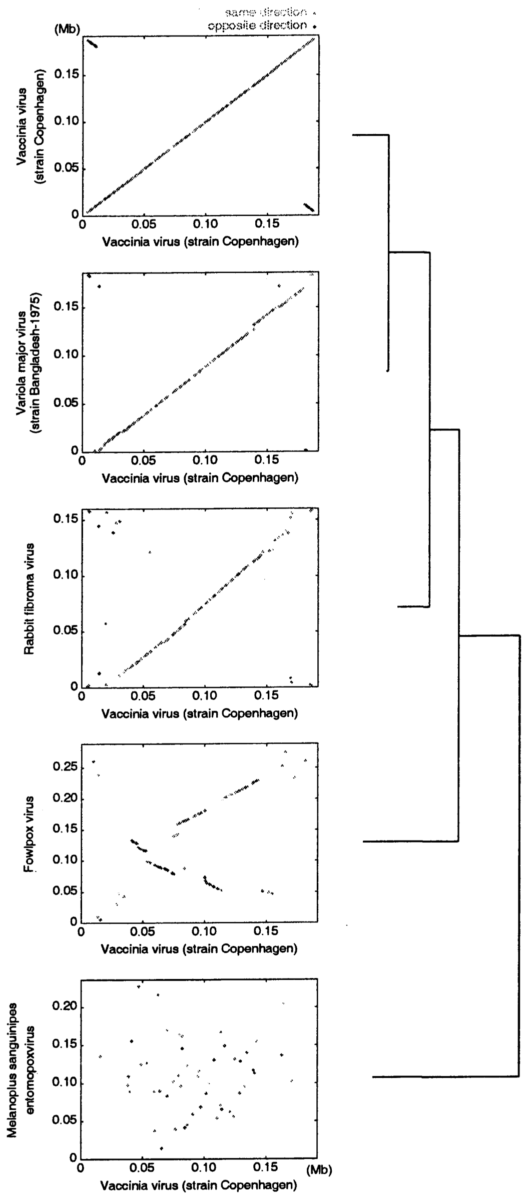

Figure 4. Gene order conservation in poxviridae. The gene order in the Vaccinia virus (strain Copenhagen) is compared with that in four other poxviridae, at varying evolutionary distances. The location of the genes in Vaccinia itself are given as a reference (figure at the top of the page). The dots indicate the position of orthologous genes in the genomes. The phylogeny at the right was constructed by clustering genomes based on the fraction of the genes they share (Snel et al., 1999).

few false positive predictions (Huynen et al., unpublished). However, the number of occurrences is very limited (Enright et al., 1999; Snel et al., 2000). This can be overcome by inferring also from paralogues (other members of multigene families with presumably different functions) at the cost of a higher fraction of false positives (Marcotte et al., 1999a).

Again, if the prediction methods are tuned so that there is a high accuracy, the increasing number of completely sequenced genomes will lead to an increase in predictive power of fusion-based method.

2.3. Gene co-occurrences

A third type of context predictions has been evolved from the observation that phylogenetic patterns of orthologs vary, i.e., the presence and the absence of proteins in the different phyla can be recorded (COG-pattern, e.g., Tatusov et al., 2000, and references therein). Furthermore, genes that co-occur in different genomes have been predicted to functionally interact (Huynen and Bork, 1998). A more detailed analysis indicated a number of successful predictions (Pellegrini et al., 1999) and this kind of context was named "phylogenetic profile". Again, although several groups work on these methods, the results differ as they very much depend on the choice of parameters and thresholds. In our hands, co-occurrence using orthology has a lower accuracy than gene fusion and gene neighborhood methods (Huynen et al., unpublished data).

2.4. Shared regulatory elements

Another context is that of shared or similar regulatory elements. Unfortunately, the identification of motifs in noncoding regions is much harder than in coding ones and the current accuracy even for known regulatory elements is in the range of 50% (see Bork, 2000, and refs. therein). Nevertheless, there have been a number of successful predictions of co-regulation and functional association of proteins based on shared regulatory elements upstream their respective genes within one organism (Hughes et al., 2000; Hertz and Stormo, 1999; Roth et al., 1998). In prokaryotes, regulatory elements seem to be able to change positions within equivalent operons in different species (Lathe et al., unpublished observation) and the functional association observed might be an indirect one. Thus, although the context via shared or similar regulatory elements has certainly potential, one has to wait for larger and validated datasets to be able to explore it in a systematic way.

2.5. Combination of methods

It has been proposed that combining several of the approaches above with knowledge databases on pathways or expression data increases the signal to noise ratio even if the initial methods are tuned to create a high noise level (Marcotte et al., 1999b). Even without the inclusion of known biological facts but just by exploiting neighborhood, fusion and co-occurrence approaches tuned to achieve a high accuracy (on the cost of less instances for predictions), the overall coverage and accuracy seems very high. It has to be noted that there is a high overlap with classical, homology-based predictions (Huynen et al., unpublished observation). One can exploit the complementarity between the methods, and thus predict both the molecular function of a protein by homology analysis and the pathway in which it plays a role by context analysis.

2.6. STRING: an implementation for the prediction of functionally associated genes

One implementation for the combination of different context methods with knowledge databases has been reported that allows a high rate of false positives in each individual method, but hopes for a drastic improvement by requiring detection of the same functional association by at least two of the methods employed (Marcotte et al., 1999b). Although the noise ratio remained relatively high, it can be successfully employed to reveal candidates that can then experimentally verified (e.g., to find additional drug targets that associate with a known one). An alternative approach is to require a high accuracy for each individual method and add the results. Starting with detecting conservation of genes within potential operons and gene fusions, and in addition considering co-occurrences of genes, is the strategy behind a web-based tool named STRING (Search Tool for Recurring Instances of Neighboring Genes; **http:www.bork.embl-heidelberg.de/STRING).** When querying the tool with a single gene of one of the currently about 30 completely sequenced genomes that are incorporated, orthologs in other genomes and all neighboring genes are retrieved. If some neighboring genes are found to be conserved, the procedure can be iterated for these conserved neighbors and their neighbors and so forth (whereby fusion and the pure presence in the respective operons are also considered). In the last iteration (the number can be specified by the user or is the one that leads to convergence of the procedure: i.e. no more context has been detected), co-occurrence is being retrieved of all genes detected this way. The iterative approach of the algorithm allows one to detect complete pathways. For example, all known genes of the tryptophan biosynthesis are retrieved when starting with a single gene of this pathway, TrpA (tryptophan synthase, alpha subunit) without any additional false positive (Figure 5).

2.7. The uber-operon: gene order conservation at a higher level

The procedure described above can reveal some surprising connections between processes previously thought to be unrelated (for example membrane synthesis enzymes and ribosomal proteins) although experimental prove is needed for such

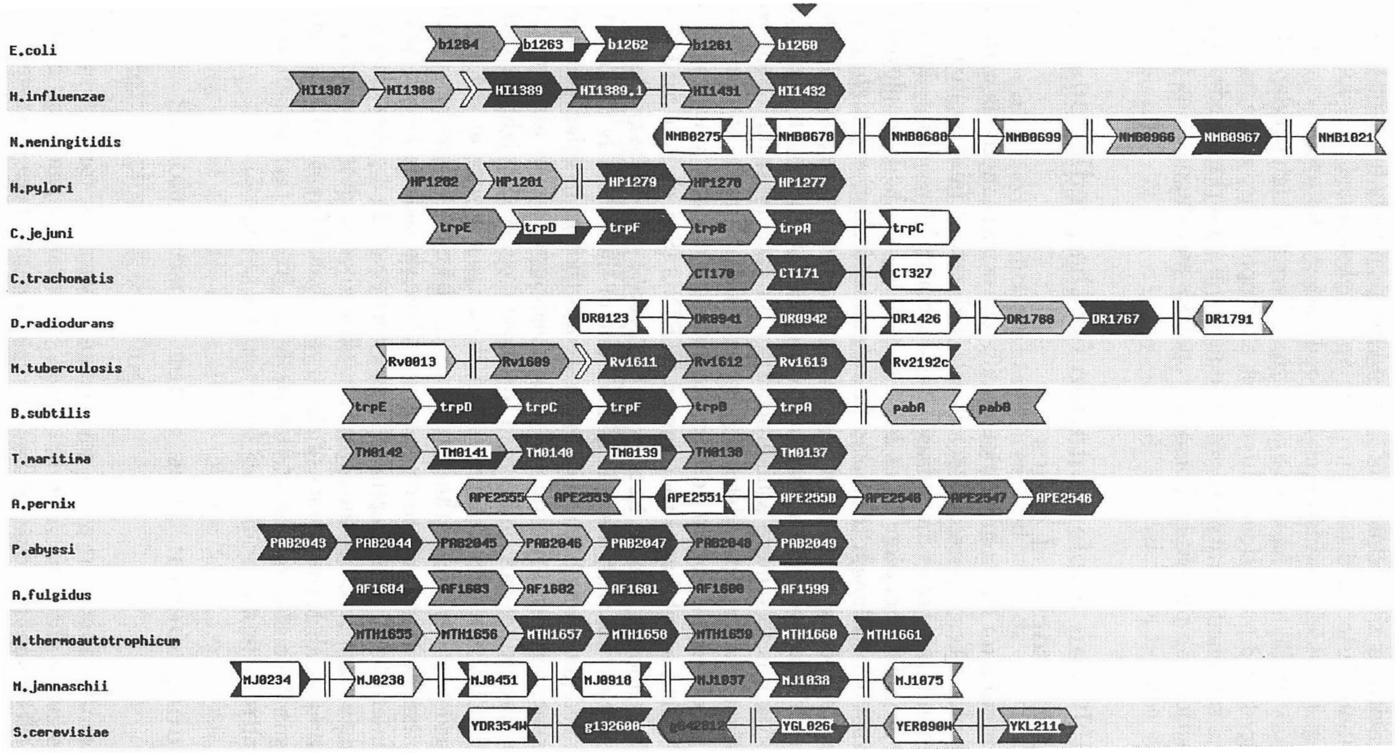

Figure 5. The genome organization of the trypthophan synthesis genes in different species (archaea, bacteria and yeast), as retrieved by STRING. The query that was used to search STRING was PAB2049, the tryptophan synthase subunit alpha from pyrococcus abbysii. For an explanation of the different symbols, see the caption of Figure 1.

predictions. It further can assign functional associations to hitherto uncharacterized proteins. Interestingly, when studying many of these examples it turns out that the gene order within operons is not conserved, but the shuffling of genes seems to be restricted. Though the genomic rearrangements vary for an individual gene's specific neighborhood, many genes are maintained over evolutionary time within the neighborhood of a discrete set of functionally related genes. We call this phenomenon an uber-operon (Figure 6).

These uber-operons can be seen as a natural classification of a cellular process. The variations we see in uber-operons, either in a single species or entire taxa, could be indicative of variations on a cellular level. Novel additions of genes in a particular species or taxa into an uber-operon might indicate new biochemical pathways or regulatory changes in that species. Also, uber-operons might be good indicators of relationships of the processes within a cell. Some uber-operons most likely share genes, the genes being in one uber-operon in one group of species and another uber-operon in a second. This could indicate the relationships and connectivity of the processes themselves. Thus, the uber-operons might form the basis for a natural classification of cellular functions and processes as well as for the characterization of novel biochemical pathways in a particular species.

Hence evolutionary constraints revealed by different types of conserved genomic context prove once more to be vital in the recognition of functional features and open a new avenue to study cellular networks.

Acknowledgements

We thank BMBF and DFG for funding and the other members of our group for fruitful discussions.

References

Aach, J., Rindone, W., and Church, G. M. 2000. Systematic management and analysis of yeast gene expression data. *Genome Research* **10:431-445.**

Andrade, M. A. and Valencia, A. 1998. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics* **14:600-607.**

Bork, P. 2000. Powers and pitfalls in sequence analysis. The 70% hurdle. *Genome Research* **10:398-400.**

Bork, P. and Koonin, E. V. 1998. Predicting function from protein sequences: Where are the bottlenecks? *Nature Genetics* **18:313-318.**

Dandekar, T., B., S., Huynen, M., and Bork, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23:324-328.**

Doolittle, R. F. 1999. Do **you dig my groove?** *Nature Genetics* **235-8.**

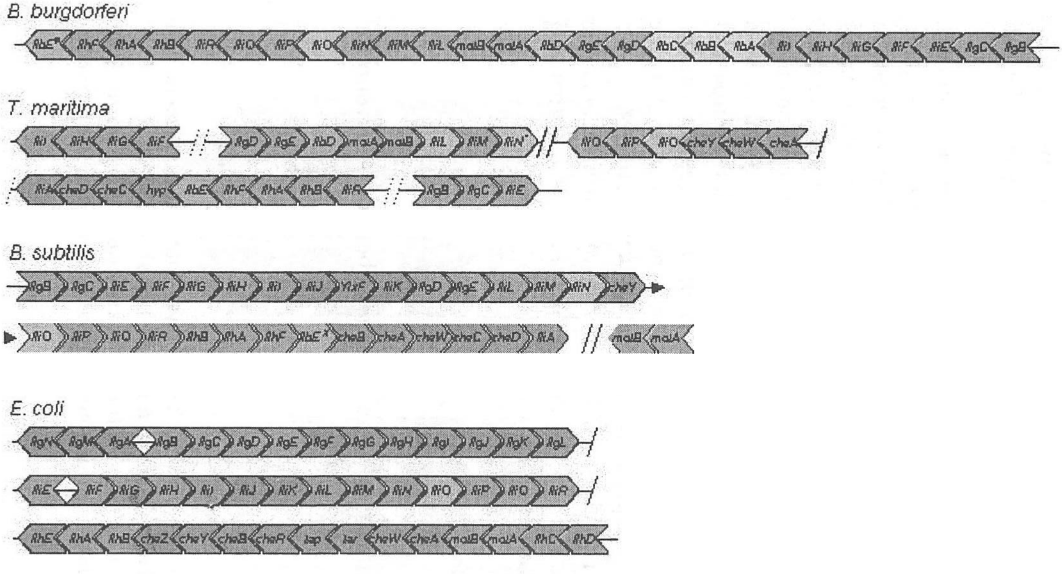
B. burgdorferi
flbE# flhF flhA flhB fliR fliQ fliP fliO fliN fliM fliL motB motA flbD flgE flgD flbC flbB flbA fliI fliH fliG fliF fliE flgC flgB
T. maritima
fliI fliH fliG fliF flgD flgE flbD motA motB fliL fliM fliN* fliQ fliP fliO cheY cheW cheA
fliA cheD cheC hyp flbE flhF flhA flhB fliR flgB flgC fliE
B. subtilis
flgB flgC fliE fliF fliG fliH fliI fliJ YlxF fliK flgD flgE fliL fliM fliN cheY
fliO fliP fliQ fliR flhB flhA flhF flbEX cheB cheA cheW cheC cheD fliA motB motA
E. coli
flgN flgM flgA flgB flgC flgD flgE flgF flgG flgH flgI flgJ flgK flgL
fliE fliF fliG fliH fliI fliJ fliK fliL fliM fliN fliO fliP fliQ fliR
flhE flhA flhB cheZ cheY cheB cheR tap tar cheW cheA motB motA flhC flhD

Figure 6. **A** single putative operon of *Borellia burgdorferi* which consists of 26 genes of the flagella structural, synthesis and taxis system is compared with related operons of the genomes of *Thermotoga maritima*, *Bacillis subtilis*, and *Escherichia coli.* Though these genes are rearranged in other genomes, they are found in the transcriptional neighborhood of other flagellar related genes. Dark blue signifies *B. burgdorferi* genes and their orthologs in each of the three genomes. Light blue genes are orthologs found in all three genomes, though genbank annotation labels the gene by a different name or as "hypothetical". Red are flagellar genes with which the orthologs of the *B. burgdorferi* operon genes have been rearranged. Yellow are *B. burgdorferi* genes where no orthologs are found in the other three genomes, while orange signifies those present in only one or two species. Gene names starting with 'f' are flagellar structural (such as motor switch, basal body, P-ring and others), synthesis or regulatory proteins. *'che'* gene names are proteins involved in chemotaxis. *'tap'* and *'tar'* in *E. coli* both code for methyl-accepting chemotaxis proteins. *motA* and *motB* encode flagella motor proteins. Notes: # the *flbE* gene from the originally described operon in *B. burgdorferi*, is a fragment of a larger ORF annotated as hypothetical. **fliN* in *T. maritima* is proceeded by an un-annotated ORF which is homologous to a domain present in the N-term of *fliN* from *B. subtilis*, the N-term of *fliM* and *cheC*. +*fliO* is annotated *as fliZ* in *B. burgdorferi*, T. *maritima*, and *B. subtilis*, though these are orthologous to *fliO* in *E. coli. fliZ* in *E. coli* has no orthologs in these three species. X *flbE* in *B. subtilis* is annotated *as* hypothetical, though it is homologous to *flbE* in *B. burgdorferi.* Gene locations and sequences for these and ribosomal genes were obtained from databases referenced at `http://www.tigr/org/tdb/`.

EISENHABER, F. AND BORK, P. 1998. Wanted: subcellular localization of proteins based on sequence. *Trends in Cell Biology* 8:169-170.

ENRIGHT, A. J. ET AL. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86-90.

GELFAND, M. S. AND KOONIN, E. V. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Research* 25:2430-2439.

HERTZ, G. Z. AND STORMO, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563-577.

HUGHES, J. D., ESTEP, P. W., TAVAZOIE, S., AND CHURCH, G. M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Evolution* 296:1205-1214.

HUYNEN, M. A. AND BORK, P. 1998. Measuring genome evolution. *Proceedings of the National Academy of Sciences USA* 95:5849-5856.

HUYNEN, M. A. ET AL. 2000. Exploitation of gene context. *Current Opinion in Structural Biology* 10. in press.

HUYNEN, M. A. AND SNEL, B. 2000. Gene and context: integrative approaches to genome analysis. *Advances in Protein Chemistry* 54:345-379.

KANEHISA, M. AND GOTO, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28:27-30.

KOLSTO, A. B. 1997. Dynamic bacterial genome organization. *Molecular Microbiology* 24.

MARCOTTE, E. M. 2000. Computational genetics: finding protein function by nonhomology methods. *Current Opinion in Structural Biology* in press.

MARCOTTE, E. M. ET AL. 1999a. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285.

MARCOTTE, E. M., PELLEGRINI, M., THOMPSON, M. J., YEATES, T.O., AND EISENBERG, D. 1999b. A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83-86.

MUSHEGIAN, *A.* R. AND KOONIN, E. V. 1996. Gene order is not conserved in bacterial evolution. *Trends in Genetics* 12:289-290.

NAKAI, *K*. 2000. Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry* 54. in press.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G. D., AND MALTSEV, N. 1999. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences USA* 96:2896-2901.

OVERBEEK, R., LARSEN, N., PUSCH, *G*. D., D'SOUZA M., SELKOV, JR., E., KYRPIDES, N., FONSTEIN, M., MALTSEV, N., AND SELKOV, E. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research* 28:123-125.

PELLEGRINI, M. ET AL. 1999. Assigning protein functions by comprative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences USA* 96:4285-4288.

REBHAN, M., CHALIFA-CASPI, V., PRILUSKY, J., AND LANCET, D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinfomatics* 14:556-664.

ROTH, F. P., HUGHES, J. D., ESTEP, P. W., AND CHURCH, *G*. M. *1998.* Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* 16:939-945.

SALI, A. *1999.* Functional links between proteins. *Nature* 402:23-24.

SANKOFF, *D*. *1993.* Analytical approaches to genomic evolution. *Biochimie* 75:409-413.

SANKOFF, *D*. AND BLANCHETTE, M. *1999.* Phylogenetic invariants for genome rearrangements. *Journal of Computational Biology* 6:431-445.

SCHERF, U. ET AL. *2000.* A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24:236-244.

SMITH, T. F. AND WATERMAN, M. S. *1981.* Identification of common molecular subsequences. *Journal of Molecular Biology* 25:195-197.

SNEL, B., BORK, P., AND HUYNEN, M. *2000.* Genome evolution: gene fusion versus gene fission. *Trends in Genetics* 16:9-11.

SNEL, B., BORK, P., AND HUYNEN, M. A. *1999.* Genome phylogeny based on gene content. *Nature Genetics* 21:108-110.

TAMAMES, J., CASARI, G., OUZOUNIS, C., AND VALENCIA, A. *1997.* Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution* 44:66-73.

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 Research* 28:33-36.

TEICHMANN, S. A. AND MITCHISON, G. *2000.* Computing protein function. *Nature Biotechnology* 18:27.

WATANABE, H. ET AL. *1997.* Genome plasticity *as* a paradigm of eubacterial evolution. *Journal of Molecular Evolution* 44:357-S64.

XENARIOS, I., RICE, D. W., SALWINSKI, L., BARON, M. K., MARCOTTE, E. M., AND EISENBERG, D. *2000.* DIP: the database of interacting proteins. *Nucleic Acids Research* 28:289-291.

EUROPEAN MOLECULAR BIOLOGY LABORATORY, MEYERHOFSTR. *1, 69117* HEIDELBERG, GERMANY
E-mail address: **bork@EMBL-Heidelberg.de**
E-mail address: **snel@EMBL-Heidelberg.de**
E-mail address: **suyama@EMBL-Heidelberg.de**
E-mail address: **dandekar@EMBL-Heidelberg.de**
E-mail address: **huynen@EMBL-Heidelberg.de**

MAX-DELBRÜCK-CENTRUM FÜR MOLEKULARE MEDIZIN (MDC) BERLIN-BUCH, *13092* BERLIN GERMANY
E-mail address: **gerrit@bioinf.mdc-berlin.de**