doi:10.1006/jmbi.2001.4701 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) **311**, 639–656

JMB

# SNAPping up Functionally Related Genes Based on Context Information: A Colinearity-free Approach

**G. Kolesov, H.-W. Mewes and D. Frishman***

*Institute for Bioinformatics*
*GSF - National Research*
*Center for Environment*
*and Health, Ingolstädter*
*Landstraße 1, 85764*
*Neueherberg, Germany*

**Corresponding author*

We describe a computational approach for finding genes that are functionally related but do not possess any noticeable sequence similarity. Our method, which we call SNAP (similarity-neighborhood approach), reveals the conservation of gene order on bacterial chromosomes based on both cross-genome comparison and context information. The novel feature of this method is that it does not rely on detection of conserved colinear gene strings. Instead, we introduce the notion of a similarity-neighborhood graph (SN-graph), which is constructed from the chains of similarity and neighborhood relationships between orthologous genes in different genomes and adjacent genes in the same genome, respectively. An SN-cycle is defined as a closed path on the SN-graph and is postulated to preferentially join functionally related gene products that participate in the same biochemical or regulatory process. We demonstrate the substantial non-randomness and functional significance of SN-cycles derived from real genome data and estimate the prediction accuracy of SNAP in assigning broad function to uncharacterized proteins. Examples of practical application of SNAP for improving the quality of genome annotation are described.

© 2001 Academic Press

*Keywords:* Genome analysis; gene function prediction; gene cluster; functional coupling; metabolic pathway

## Introduction

Computer-assisted functional assignment of gene products traditionally involves identifying a significant resemblance to an experimentally characterized protein or sequence motif. Due to the constant improvement of the sequence comparison techniques, reliable recognition of extremely distant relationships between proteins has become possible. At the same time, further progress in this direction is becoming increasingly difficult, following the rule of diminishing returns: improvements of ever smaller significance require ever growing effort and sophistication. Consequently, the quest to develop complementary, similarity-free computational approaches to elucidate gene function has been triggered. For example, methods based on the linguistic analysis of textual sequence annotation and scientific literature[1] and correlating protein amino acid composition with enzyme nomenclature have been explored.[2]

Comparative genomics shows great promise in this respect. Successful attempts have been made to correlate gene properties based on their coordinated occurrence in different genomes, similarity of the mRNA expression patterns, and patterns of domain fusion.[3,4] The availability of complete genome sequences has also made it possible to study the properties of gene products in reference to their location on the chromosome. In particular, much attention has been paid to the important feature of bacterial genomes: the presence of operons, or groups of genes that are transcribed as a unit and typically code for proteins involved in the same biochemical process. For example, over 2500 operons are presumed to exist in *Escherichia coli*, with roughly a quarter of them containing two or more genes.[5] The non-random proximity of genes involved in operons represents a specific complementary information signal not recognizable by sequence comparison.

Attempts to utilize gene order for improving genome annotation were quickly undertaken following publication of the first complete genomic

Abbreviations used: SNAP, similarity-neighbourhood approach; SN, similarity-neighbourhood; S-relationship, genes related by similarity; N-relationship, genes related by neighbourhood; SN-relationship, genes related by both similarity and neighbourhood.

E-mail address of the corresponding author: d.frishman@gsf.de

© 2001 Academic Press

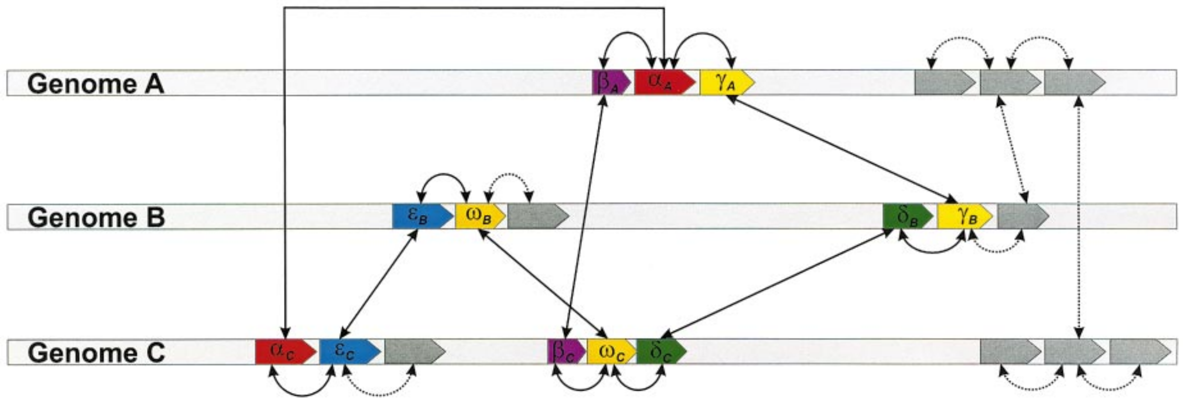

**Figure 1.** Finding genes functionally coupled with the gene $\alpha$ residing in the genome A. Colored arrows represent individual genes and their direction. Straight black arrows represent S-relationships between orthologs in different genomes, while round black arrows represent N-relationships between genes in the same genome. Only one neighbor of every gene in each direction is considered. The analysis starts with finding neigbours of gene $\alpha$, genes $\beta_A$ and $\gamma_{A,}$ in the genome A. Then their orthologs on other genomes are identified, and so on. As a result, a chain of alternating similarity and neighborhood-relationships, called an SN-graph, is constructed. In this example, the SN-graph has a closed path $\alpha_A,\gamma_{A,}\gamma_{B,}\delta_B,\delta_C,\omega_C,\omega_B,\varepsilon_B,\varepsilon_C,\alpha_C,\alpha_A$, or SN-cycle, indicating that at least some part of the constituent genes may be functionally related. Continuous black arrows correspond to the closed path while the rest of the SN-graph is shown in broken arrows. Genes not participating in the closed path are shown in grey.

sequences, but the initial enthusiasm was dissipated by the finding that gene order is, in general, poorly conserved among phylogenetically distant species.[6–8] Hence, we are confronted with two phenomena: on the one hand, a vast number of genes from individual genomes form well-defined gene clusters; on the other hand, the conservation of these gene clusters between distant species is quite poor. However, while there is no long-range colinearity between functionally related genes in bacterial genomes, short conserved strings of genes, often confined to just two elements, appear to be relatively widespread.[9,10] This observation has recently been exploited to create a similarity-free algorithm for gene function prediction.[11,12] If a pair of genes, A and B, is spatially conserved across many or all completely sequenced genomes, the chances are that they belong to the same gene cluster, that their functions are related, that they may physically interact with one another,[6,13] and that they are possibly co-regulated. Thus, if the function of the product B is not known, sustained spatial proximity to gene A of known function may help to draw conclusions about its cellular role, even if no similarity to known proteins could be detected. Moreover, even if the function of gene A is not known, knowledge about the adjacency of A and B may prove useful, e.g. in large-scale functional analysis strategies and for finding drug targets.[13] This approach was dubbed ''guilt-by-association'' because of its obvious relation to common investigative practices: persons often seen in the company of known criminals may become suspects themselves. Can we go further in applying such detective methods? What if a middleman is involved?

In the work of Overbeek *et al.*[12] functional coupling of genes on the chromosome was deduced based on short-range colinearity between genes. The goal of this work was to uncover functional coupling between co-regulated proteins in prokaryotic genomes using the conservation of gene order beyond mere colinear gene clusters and to detect loosely coupled genes, not necessarily residing in adjacent chromosomal locations. This is achieved by identifying orthologs of a given gene in other genomes, considering their neighbors, finding the orthologs of these neighbors, and so on. Continuing the underworld analogy, this would be equivalent to linking two mafia bosses who were never seen together through a chain of their interacting subordinates. We show that the chains of alternating similarity and neighborhood relationships between genes in multiple genomes are strongly non-random and very informative for annotating functionally uncharacterized genes.

Following the established bioinformatics tradition, we tried very hard to invent a conspicuous acronym to name our technique. For the lack of a better idea, we dub it SNAP, for similarity-neighborhood approach.

## Main ideas and definitions

Genes fulfilling the same function in different organisms (orthologs), or similar but distinct functions in the same organism (paralogs) are expected to possess a certain degree of sequence similarity due to the evolutionary conservation of their primary structure. By contrast, functionally related genes are essentially different genes that are involved, for example, in the same metabolic or signaling pathway. Such genes are normally not similar; hence, their relatedness is not detectable by sequence comparison. Instead, functionally related genes often form clusters on the chromosome;[14] their relatedness may be manifested by spatial proximity rather than structural resemblance. Throughout this text, we will use the terms S-relationship, N-relationship, and SN-relationship to describe the cases where genes are related by similarity, neighborhood, or a mixture thereof, respectively.

In this work, we attempt to exploit the observation that neighboring genes on bacterial chromosomes tend to be functionally related, even if there is no evidence that their positional preference with respect to each other is conserved across many different genomes. Potentially, any random pair of adjacent genes could be functionally coupled. It is evident, of course, that many hundreds and even thousands of genes encoded in complete bacterial genomes fall into hundreds of different functional categories, making the joint occurrence of two functionally related genes a rather unlikely event.[15] We need to be able to distinguish random pairs of physically proximate genes from meaningful ones, without relying, in general, on the conservation of such pairs across multiple genomes.

Before we provide a formal description of our algorithm (see Materials and Methods), we start with a simple illustration. Let us first consider a group of five genes involved in a certain biochemical process, and compare this group as a whole with functionally related groups in other genomes. In the case of a perfectly conserved gene cluster, we will observe a string of genes $\alpha_A,\beta_A,\gamma_{A,}\omega_A,\varepsilon_A$ in the genome A, $\alpha_B,\beta_B,\gamma_{B,}\omega_B,\varepsilon_B$ in the genome B, $\alpha_C,\beta_C,\gamma_{C,}\omega_C,\varepsilon_C$ in the genome C, and so on, such that the genes from different genomes denoted with the same Greek letter are S-related, and the genes from the same genome are N-related. In a more complex, and more realistic case, many of the inter-genome S-relationships may not be preserved due to physiological differences between the species involved, or simply because the similarity is not detectable with current sequence comparison tools. Likewise, and even more probably, the N-relationships within each genome may be disrupted as a result of gene shuffling in the course of evolution. Therefore, the association between the different instances of this particular gene cluster in different genomes will be expressed as an irregular mixture of S and N-relationships.

Let us consider a hypothetical example, depicted in Figure 1, and focus on the chain of SN-relationships originating from gene $\alpha$ in genome A. This gene is N-related to the genes $\beta_A$ and $\gamma_A$. Gene $\gamma_A$ is S-related to $\gamma_B$, the latter is N-related to $\delta_B$, and so on. The complete system of such SN-relationships, subject to certain limitations described below, forms an SN-graph. SN-paths on the graph are made up of alternating S and N-relationships. The former are derived using selective sequence comparison tools, such as BLAST,[16] and are thus extremely significant. By contrast, the latter are overwhelmingly random. For this reason, the majority of the SN-paths has no diagnostic value. However, intermixed with a large number of "false positives" among N-relationships, i.e. pairs of totally unrelated genes, are a number of N-related genes that are actually functionally coupled. We put forward a hypothesis that such meaningful N-relationships are likely to occur in closed SN-paths, which we will call SN-cycles. In Figure 1, the longest SN-cycle is represented by the path $\alpha_A,\gamma_{A,}\gamma_{B,}\delta_B,\delta_C,\omega_C,\omega_B,\varepsilon_B,\varepsilon_{C,}\alpha_C,\alpha_{A.}$ The primary intuition here is that the N-relationships resulting from non-random associations between genes will have a statistical tendency to throw a bridge between pairs of S-related proteins, and ultimately help join proteins that belong to the same metabolic pathway, resulting in a closed path on the graph. Our principal approach in this work is to exploit simultaneously the two possible types of relatedness between genes (S and N-relationships) in order to establish functional links undetectable by either type of relationship alone.

## Results and Discussion

### Formal properties of SN-cycles

We begin with asking two questions: (i) do non-trivial SN-cycles (i.e. those not involving colinear gene clusters) exist; and (ii) if they exist, what is the chance that they occur at random. To answer the first question, it is sufficient to provide an example. Figure 2 shows a closed system of SN-relationships involving some of the genes responsible for lysine biosynthesis in five prokaryotes. There are three adjoining SN-cycles originating at the *E. coli* gene coding for dihydrodipicolinate reductase. The detailed discussion of this example from the functional point of view will follow later.

In order to answer the second question, we have studied the behavior of SN-graphs and their dependence on various analysis parameters using a set of 12 completely sequenced genomes from phylogenetically distant species (see Materials and Methods). Figure 3(a) shows the dependence of the number of SN-cycles identified from the number of genomes used in the analysis. The graph makes immediately obvious the value of a large number of sequenced genomes in comparative genomics: there is a boost in the number of SN-cycles found as the number of genomes approaches ten. This is in agreement with the results of Overbeek *et al.*, who noted that in order to detect functional coupling for a given functional subsystem, at least ten genomes are needed.[12]

The same experiment was performed with our set of 12 genomes after randomly shuffling the gene order within each genome, which effectively leads to destroying meaningful N-relationships while keeping S-relationships intact. The difference in the occurrence of SN-cycles in real and shuffled genomes quickly grows with the number of genomes and becomes especially pronounced when more than ten genomes are considered. In the com-

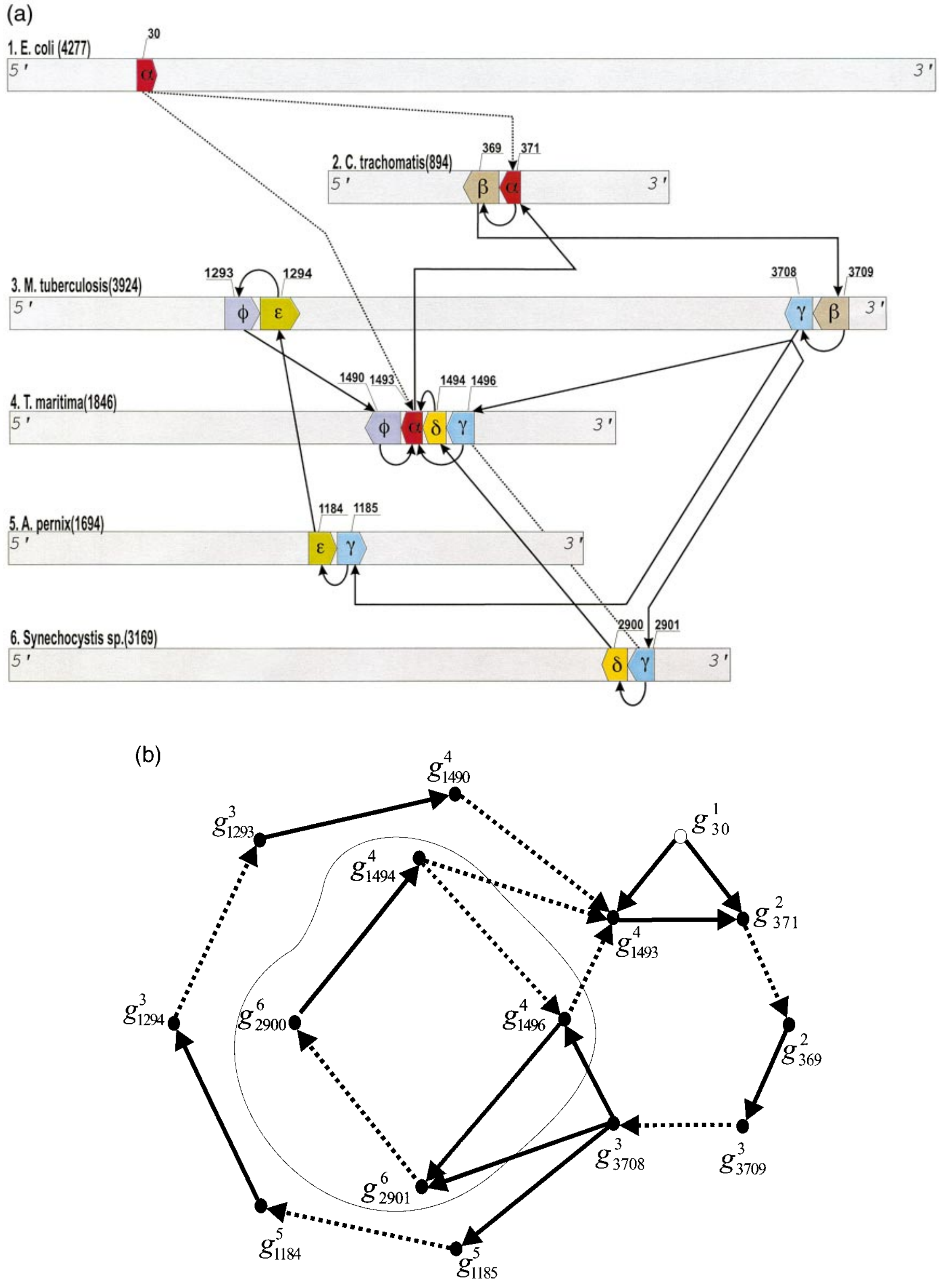

**Figure 2**. (*legend opposite*).

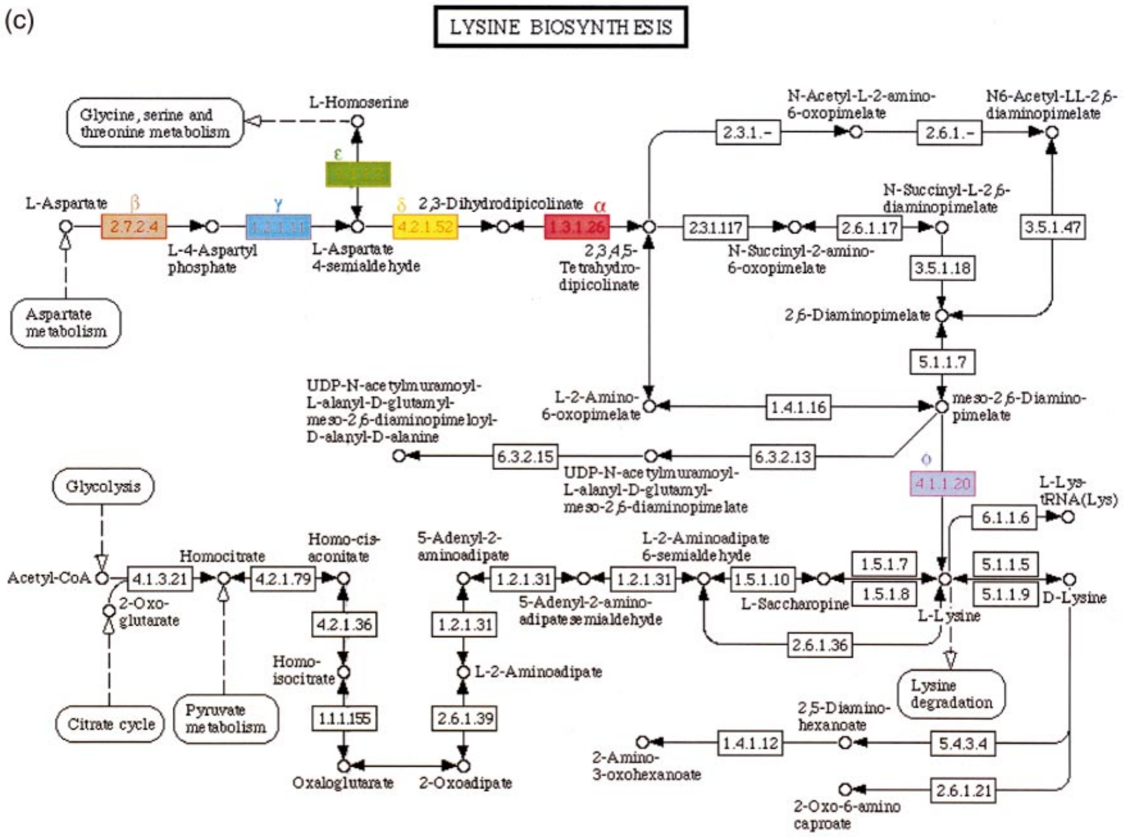

**Figure 2.** SNAP analysis of the *E. coli* gene g1786214 coding for dihydrodipicolinate reductase. A part of the SN-system originating from the *Chlanydla trachomatis* gene gi_3328787 (which is orthologous to the *E. coli* gene) is shown. For illustration purposes, only six prokaryotic genomes are considered, numbered from 1 to 6. (a). A representation of the gene location and their S and N- relationships. The total number of genes in each genome is shown in parentheses. Sequential numbers of genes, counting from the 5′ to the 3′ end of the genome are indicated. Additionally, each gene is colored and labeled with a Greek letter according to its function: α (red), dihydrodipicolinate reductase (EC 1.3.1.26); β (brown), aspartokinase (EC 2.7.2.4); γ (cyan), aspartate-semialdehyde dehydrogenase (1.2.1.11); δ (yellow), dihydrodipicolinate synthase (EC 4.2.1.52); ε (green), homoserine dehydrogenase (EC 1.1.1.3); ϕ (lilac), diaminopimelate decarboxylase (EC 4.1.1.20). Three adjoining SN-cycles are present: (i) $g^2_{371}$ $g^2_{369}$ $g^3_{3709}$ $g^3_{3708}$ $g^4_{1496}$ $g^4_{1493}$;(ii) $g^2_{371}$ $g^2_{369}$ $g^3_{3709}$ $g^3_{3708}$ $g^5_{1185}$ $g^5_{1184}$ $g^3_{1294}$ $g^3_{1293}$ $g^4_{1490}$ $g^4_{1493}$; and (iii) $g^2_{371}$ $g^2_{369}$ $g^3_{3709}$ $g^3_{3708}$ $g^3_{3708}$ $g^6_{2901}$ $g^6_{2900}$ $g^4_{1494}$ $g^4_{1493}$. Incidentally, a simple colinear gene cluster involving the spatially conserved pair of genes β and γ in *T. maritima* and *Synechocystis* sp. is present; the extra S-relationship between the genes of the type γ is shown as a broken line. (b) An SN-graph corresponding to the system shown in (a). The shadowed part of the graph stems from the conserved pair of adjacent genes that have sequential numbers 1494 and 1496 in the genome of *T. maritima* and number 2900 and 2901 in the genome of *Synechosistis* sp. (c) A part of the KEGG metabolic map involving the six genes predicted to be functionally coupled. Enzymes (highlighted in the same colors as used in (a)) encoded by the genes α, β, γ, δ and ε catalyze subsequent reactions in the lysine biosynthesis pathway, while the reaction catalyzed by the enzyme ϕ is separated from the nearest reaction of the first group by two other metabolic steps.

plete set of 12 genomes with real gene order, 33,000 SN-cycles were found, as opposed to 3500 SN-cycles in shuffled genomes. It should also be noted that at greater evolutionary distances between species, the share of non-random SN-cycles increases. We thus estimate that with a sufficiently large number of evolutionary distant genomes taken into account, approximately 90% of SN-cycles are non-random. Moreover, as seen in Figure 3(b), the increase in the number of SN-cycles is almost exclusively caused by long (more than ten nodes) SN-cycles. Due to the virtual disappearance of long SN-cycles after shuffling, we are compelled to conclude that the majority of all such cycles reflect conserved spatial association between genes, although certain parts of these cycles may still be random.

As expected, detection of SN-cycles is strongly influenced by the choice of the BLAST alignment parameters (Figure 3(c) and (d)); their number grows quickly as the BLAST parameters are changed from very stringent (*E*-values close to 0, coverage close to 100%) to entirely permissive (any *E*-value, any coverage). However, even with the most permissive parameters, the number of SN-cycles identified in real, unshuffled genomes is nearly an order of magnitude higher than in the genomes with random gene order. Since the

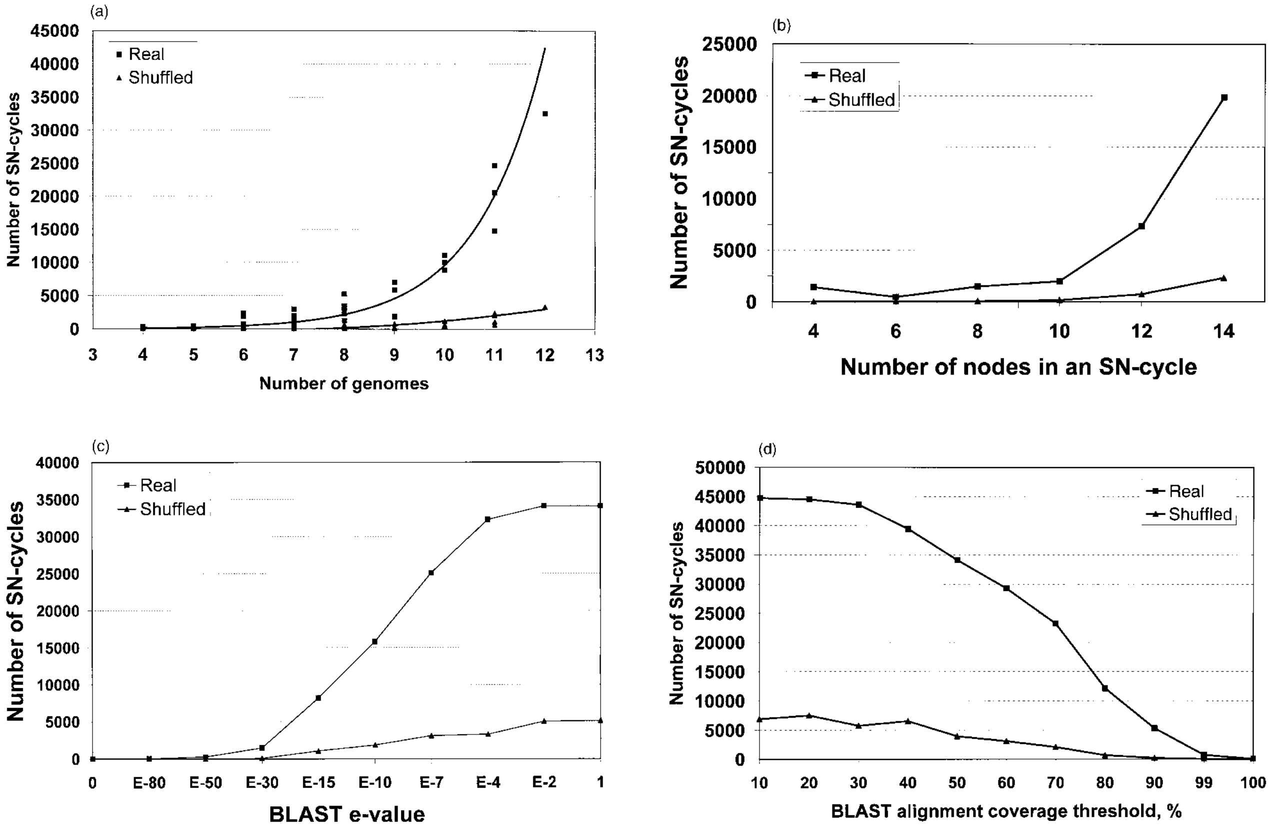

**Figure 3.** Comparison of the global properties of SN-cycles in real (squares) and shuffled (triangles) genomes. Dependence of the number of SN-cycles detected on (a) the number of genomes considered (in order to make computations feasible, only selected data points were computed), (b) cycle length, (c) BLAST cutoff *E*-value, and (d) BLAST alignment coverage is shown. The default parameters, unless explicitly specified are: BLAST cutoff *E*-value, 0.0001; BLAST coverage, 0.4; number of genomes, 12.

S-relationships are not influenced by gene order shuffling, the difference observed is solely due to the strong functional coupling of adjacent genes in the former and the virtual disappearance of the N-relationships in the latter.

## Functional content of SN-cycles

Now that we have formally established the overwhelming non-randomness of long SN-cycles and their frequent occurrence, it is time to examine their functional content. The central issue in accessing the performance of our method is the granularity of the functional assignments. Similarity-free approaches are necessarily less specific than methods based on protein sequence and structure comparison. While the latter are often capable of predicting precise specificity of a certain enzyme, the former are intended to attribute proteins to broad functional classes or predict their involvement in the same physiological processes or cellular structures.

Let us consider again the example shown in Figure 2. The system of three adjoining SN-cycles links six different enzymes participating in the lysine biosynthesis pathway (Table 1). As seen in Figure 2(c), five of these proteins ($\alpha$,$\beta$,$\gamma$,$\delta$ and $\varepsilon$) catalyze subsequent reactions, while the reaction catalyzed by the enzyme $\phi$ is separated from the nearest reaction of the first group by two intervening steps, corresponding to a metabolic distance $D = 3$. Assuming normalization coefficient $\lambda_p = 1$, the pathway coefficient (see Materials and Methods) will be equal $K_p = 1(5/6) \approx 0.83$ for $D_t = 1$, and $K_p = 1$ for $D_t \geqslant 3$. Further, all six proteins belong to the same functional role category 01.01.01 (amino acid biosynthesis), which means that the functional category coefficient in this case will be $K_f = 1(6/6) = 1$ (again, assuming $\lambda_f = 1$). Thus, both coefficients indicate a high degree of functional coupling between the enzymes considered. Importantly, none of these three SN-cycles or their parts constitutes a conserved colinear gene cluster, although one such cluster is incidentally present and involves the conserved pair of genes coding for dihydrodipicolinate synthase and homoserine dehydrogenase shared between the *Thermotoga maritima* and *Synechocystis* sp. genomes.

To assess the global performance of our method, we have studied the behavior of the $K_p$ measure on the full set of SN-cycles delineated from 12 genomes. The complete KEGG pathway database[17] was treated as a set of separate subgraphs corresponding to the individual biochemical pathways, such as lysine biosynthesis or glycolysis. Effectively, by using such an approach we are introducing additional *a priori* knowledge about functionally coupled genes in our measurements. Using this approach (Figure 4(a)) to estimate $K_p$ leads to a good separation between real and shuffled genomes for all values of the maximally allowed metabolic distance *D*: the functional content of realistic SN-cycles appears to be an order of magnitude higher. Such bias would not have any influence on $K_p$ if gene groups found by SN-cycles were random.

Comparison of SN-cycles in real and shuffled genomes in terms of the pathway coefficient $K_p$ is presented in Figure 4(b). Over 30 % of all real SN-cycles found have $K_p$ values greater than 0.5, in contrast to only 1 % of random cycles. Even in the range $0.2 < K_p < 0.5$, real SN-cycles have a nearly fivefold lead over the random ones, and the total of 81 % of the cycles are in the range $0.2 < K_p < 1.0$. By contrast, the same comparison for the functional category coefficient $K_f$ (Figure 4(c)) shows that only 40 % of the real SN-cycles are in the range $0.2 < K_f < 1.0$, while 60 % have lower $K_f$ values and cannot be statistically distinguished from random cycles. We can thus conclude that the SNAP algorithm is capable of associating gene products involved in a common biochemical pathway, while the specific functions of individual genes represented in terms of a cellular role category appear to be correlated rather weakly.

## Estimating the predictive power of SNAP

The following simple considerations provide the basis for the estimation of the predictive power of SNAP. Suppose a gene of interest is grouped in an SN-cycle together with a number of other genes with known EC numbers and an arbitrary number of genes without EC numbers assigned. We will ignore the latter, since they make no contribution to the automatic annotation of the query gene. Assuming that at least one gene with a known EC number is related to the query gene, the probability of a correct functional coupling prediction for these particular query gene and SN-cycle is equal to the pathway coefficient $K_p$ of the cycle. However, it may happen that none of the genes in the SN-cycle is pathway-related to the query sequence. Thus, the expected probability of a correct prediction for a given SN-cycle should, on average, be somewhat lower than its $K_p$, dependent on the frequency of occurrence of a particular functional class. For each gene characterized through SNAP, we calculated $K_p$ of the SN-cycle used for the prediction and compared the pathway assignment of the most represented gene group in the cycle with that of the query gene. Two alternative conditions for considering a prediction of functional coupling to be correct were utilized: (a) best group condition, when the query gene was found in the same pathway as the genes of the single most represented enzyme group in all of the cycles associated with the query gene; and (b) all groups condition, when the query gene was found in the same pathway as the genes of any enzyme group across all cycles.

The cumulative graph in Figure 5 shows the dependence of the SNAP best group prediction accuracy on the minimal allowed $K_p$ coefficient based on our data. The average success rate for the entire set of genes participating in the SN-cycle is around 45 %. If one considers only SN-cycles with $K_p > 0.4$, the prediction accuracy increases to over 75 %. As seen in Figure 4(b), approximately 60 % of all SN-cycles in real genomes (as opposed to only 7 % in shuffled genomes) have the $K_p$ coefficient in this range. Not surprisingly, the percentage of true positives for the shuffled genomes shown in Figure 5 remains constant for all values of $K_p$. Note that the curve for real SN-cycles in Figure 5 tails off somewhat at $K_p$ values greater than 0.9. This happens because many of the SN-cycles with $K_p$ values equal to exactly 1.0 include only two genes with known EC numbers, while SN-cycles with $K_p$ values in the range 0.8-1.0 are typically calculated on the basis of five to ten genes (data not shown). The probability of encountering two out of two genes with the same EC number by chance is higher than, for example, to find eight out of ten genes with the same EC number. In other words, this curve is not normalized by the number of genes actually used to calculate $K_p$.

In Table 2 we present the percentage of true positive predictions for the individual genomes studied measured as described above. Only SN-cycles with $K_p$ greater than 0.4 were considered. The best group true positive rate for such cycles varies from 54 % for *Mycoplasma pneumoniae* to 90 % for *Synechosystis* sp., while the all groups numbers lie in the range from 63 % (*M. pneumoniae, Treponema pallidum*) to 91 % (*Campylobacter jejeuni*). Overall, the all groups true positive rate is somewhat better than the best group simply because the odds of finding genes coupled with the query gene in many KEGG pathway maps are higher than in just one map.

## Genome annotation with SNAP

The genome of the thermoacidophilic archaeon *Thermoplasma acidophilum* containing 1507 predicted genes has recently been sequenced and subjected to careful manual annotation† using the PEDANT software system.[28] In particular, each gene was assigned to one of the following categories, reflecting the current level of knowledge about its biochemical function: known protein (24 genes); strong similarity to known protein (189 genes); similarity to known protein (495 genes); weak similarity to known protein (101 genes); strong similarity to unknown protein (110 genes);

† Available online at http://pedant.gsf.de/cgi-bin/wwwfly.pl?Set = Tacidophilum&Page = index

**Table 1.** Genes constituting the SN-cycle shown in Figure 4 (shadowed) and their orthologs

| Genome | Gene | ID | Description | Start | Stop |
|---|---|---|---|---|---|
| *E. coli* | α | g1786214 | Dihydrodipicolinate reductase | 28374 | 29195 |
| | β | g1790455 | Lysine-sensitive aspartokinase III | 4230812 | 4229463 |
| | γ | g1788658 | usg-1 protein | 2434669 | 2433656 |
| | δ | g1788823 | Dihydrodipicolinate synthase | 2597780 | 2596902 |
| | ε | g1786183 | Aspartokinase I/homoserine dehydrogenase | 337 | 2799 |
| | ϕ | g1789203 | Diaminopimelate decarboxylase | 2976921 | 2975659 |
| *C. trachomatis* | α | gi_3328787 | Dihydrodipicolinate reductase | 415997 | 415236 |
| | β | gi_3328785 | Aspartokinase III | 414229 | 412934 |
| | γ | - | - | - | - |
| | δ | gi_3328784 | Dihydrodipicolinate synthase | 412923 | 412063 |
| | ε | - | - | - | - |
| | ϕ | - | - | - | - |
| *M. tuberculosis* | α | rv2773c | dapB dihydrodipicolinate reductase | 3082337 | 3081600 |
| | β | rv3709c | ask aspartokinase | 4153480 | 4152215 |
| | γ | rv3708c | asd aspartate semialdehyde dehydrogenase | 4152214 | 4151177 |
| | δ | rv2753c | dapA dihydrodipicolinate synthase | 3067120 | 3066218 |
| | ε | rv1294 | thrA homoserine dehydrogenase | 1449373 | 1450698 |
| | ϕ | rv1293 | lysA diaminopimelate decarboxylase | 1448026 | 1449369 |
| *T. maritima* | α | gi_4982086 | Dihydrodipicolinate reductase | 1516426 | 1516426 |
| | β | gi_4982084 | Aspartokinase II | 1515057 | 1513852 |
| | γ | gi_4982089 | Aspartate-semialdehyde dehydrogenase | 1518990 | 1518007 |
| | δ | gi_4982087 | Dihydrodipicolinate synthase | 1517307 | 1516423 |
| | ε | gi_4981061 | Aspartokinase II | 574428 | 572209 |
| | ϕ | gi_4982083 | Diaminopimelate decarboxylase | 1513842 | 1512682 |
| *A. pernix* | α | - | - | - | - |
| | β | gi_5104810 | 473 residue hypothetical aspartate kinase | 711805 | 713226 |
| | γ | gi_5104813 | Long hypothetical aspartate-semialdehyde dehydrogenase | 713223 | 714272 |
| | δ | - | - | - | - |
| | ε | gi_5104814 | Long hypothetical homoserine dehydrogenase | 714263 | 715267 |
| | ϕ | - | - | - | - |
| *Synechocystis* sp. | α | gi_1651716 | Dihydrodipicolinate reductase | 77406 | 77406 |
| | β | gi_1653765 | Aspartate kinase | 3333243 | 3335045 |
| | γ | gi_1001379 | Aspartate beta-semialdehyde dehydrogenese | 3248483 | 3249325 |
| | δ | gi_1001380 | Dihydrodipicolinate synthase | 3249385 | 3250290 |
| | ε | gi_1001182 | Homoserine dehydrogenase | 2627873 | 2626572 |

similarity to unknown protein (265 genes); weak similarity to unknown protein (85 genes); no similarity (237 genes); and questionable ORF (one gene).

Here, we focus on the 460 *T. acidophilum* genes, or roughly 30 % of the gene complement, that possess some degree of similarity to uncharacterized proteins. The number of genes of this type for which a SNAP prediction can be made depends critically on the number of genomes considered and reaches 140, or roughly one-third of this pool, when all 12 genomes are taken into account. This number will definitely grow as more genomes are included in the analysis. It appears that with a sufficient number of phylogenetically distant genomes available, essentially every gene in a genome under scrutiny will participate in at least one SN-cycle.

Let us consider the SNAP results for the *T. acidophilum* gene Ta0740. This gene, described as conserved hypothetical protein, has orthologs in a number of other bacterial genomes, but all of them

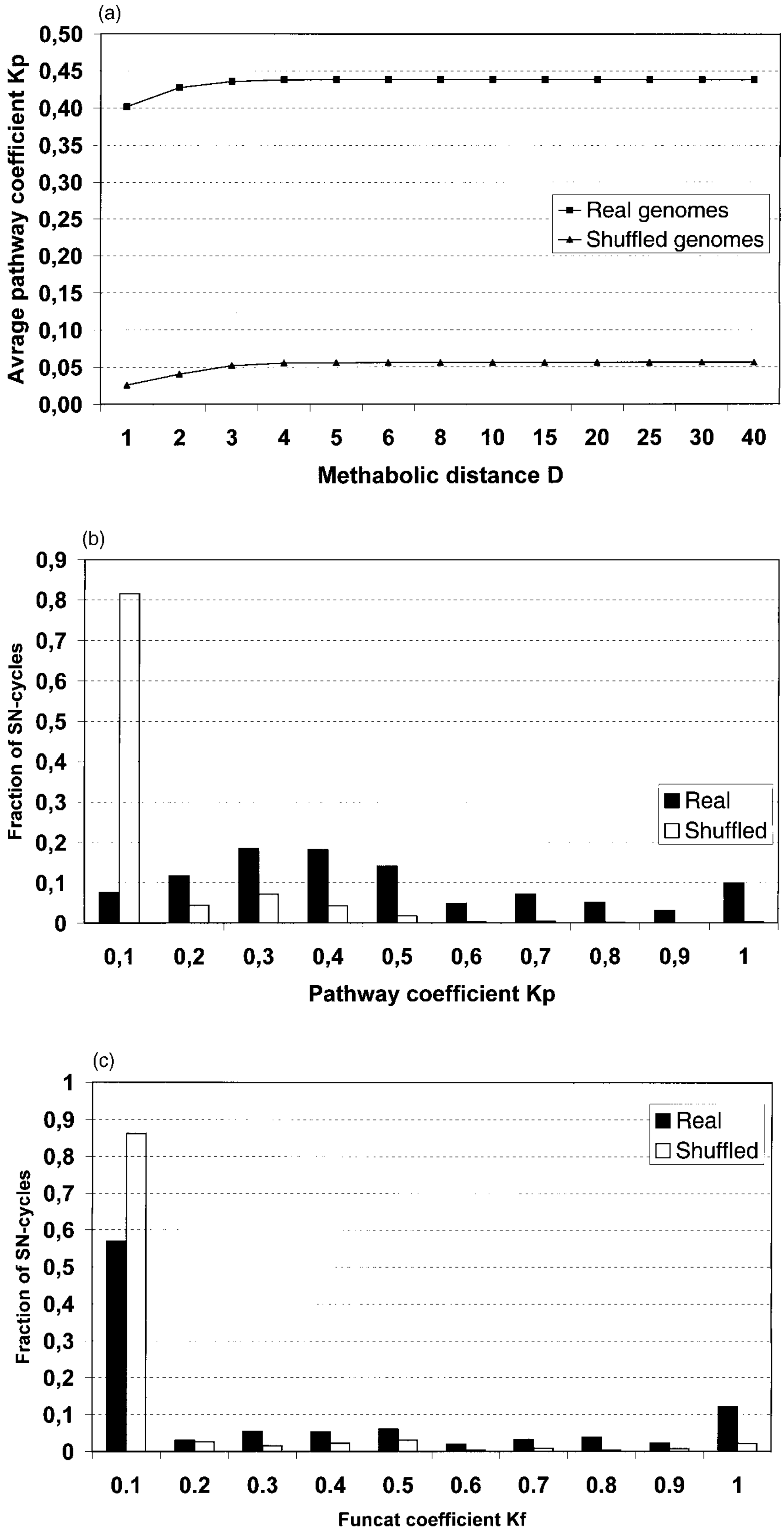

**Figure 4.** Functional content of SN-cycles in real (squares, filled bars) and shuffled ( triangles, open bars) genomes. (a) Dependence of the pathway coefficient $K_p$ on the maximal allowed metabolic distance $D$. (b) Relative occurrence of SN-cycles with different $K_p$ values. (c) Relative occurrence of SN-cycles with different values of the funcat coefficient $K_f$.

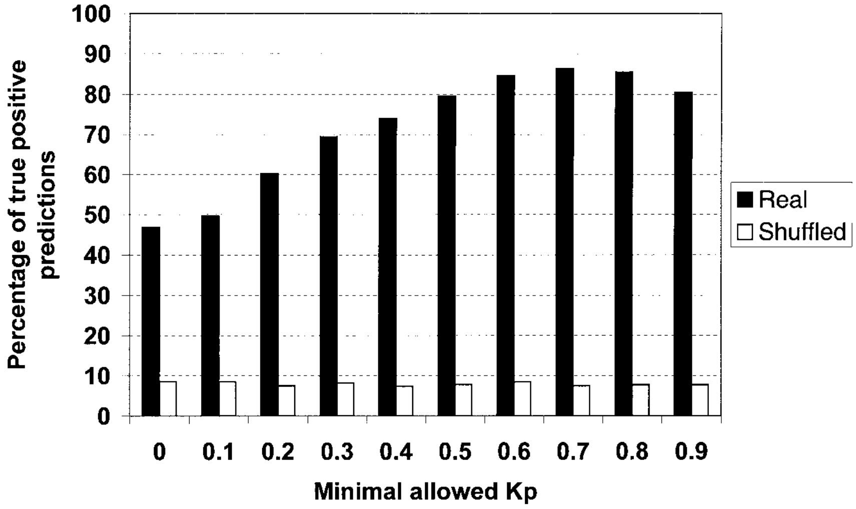

**Figure 5.** Dependence of the percentage of true positive SNAP predictions from the minimal allowed pathway coefficient $K_p$ for real (filled bars) and shuffled (open bars) genomes.

are functionally uncharacterized. The SN-cycle associated with Ta0740 (denoted α, see Figure 6(a)) involves six other types of proteins. Five of them (β, δ, ε, ζ and η) are enzymes with known EC numbers, while the sixth protein, denoted γ, is annotated as chloroplast import-associated channel IAP75. Using our software, we were able to establish that four of the enzymes, δ, ε, ζ and η, catalyze a compact group of biochemical reactions in the phenylalanine, tyrosine, and tryptophan biosynthesis pathway (KEGG map 00400, see Figure 6(b)), while the enzyme β and the non-enzymatic protein γ are seemingly unrelated to the first four proteins. Thus, based on these automatically derived KEGG assignments, the value of $K_p$ for this particular SN-cycle is $4/5 = 0.8$, because four out of five proteins with known EC numbers belong to the same metabolic pathway. However, by additional manual analysis we were able to find out that the enzyme β, involved in purine methabolism (KEGG map 00230), is actually only six reactions away from the enzyme ε. Moreover, even the protein γ with no apparent enzymatic activity may be linked to the photosynthesis system that is adjacent to the KEGG map presented in Figure 6(b) (see upper left corner). Based on the SNAP results, we predict that Ta0740 is involved in phenylalanine, tyrosine, and tryptophan biosynthesis.

The second example from *T. acidophilum* is a SNAP prediction for the gene Ta0420 (Figure 7). In the current annotation, this gene is described as conserved hypothetical protein and has similarity to hypothetical proteins in *Methanobacteirum thermoautotrophicum* and *E. coli*. Based on the comparison with the eukaryotic genome of *Saccharomyces cerevisiae*, functional categories regulation of carbohydrate utilization, other energy generation activities and carbohydrate utilization were assigned automatically by the PEDANT system to this protein; these assignments, however, are based on quite weak similarities and are thus questionable. SNAP detected two SN-cycles: a short four-node cycle composed of the proteins of α and β types, and a long cycle involving the genes α, β, γ, δ, ε, ζ and η (Figure 7(a)). The first cycle represents the case of a weakly conserved colinear gene pair: the genes α and β appear in close proximity in just two relatively close genomes (*M. thermoautotrophicum* and *T. acidophilum*). Consequently, based on the annotation of the gene β, we can putatively assign function to the gene α. Specifically, functional categories automatically assigned to β by PEDANT do indeed coincide with those assigned to α (see above) and thus confirm them (Figure 7(b)).

The long SN-cycle reveals the following: α, β, γ and ζ were assigned to the functional category carbohydrate utilization (β, γ and ζ are well-known enzymes occurring in the glycolysis pathway and other energy-related pathways), gene η is a regulatory protein of unclear function, gene δ is a carbonic anhydrase (whose functional role is also not clear) and gene ε is described as NifU-related protein (Figure 7(b)). NifU protein is involved in the nitrogen fixation process in certain soil bacteria and cyanobacteria. In our example, though, it has orthologs in *Chlamydia pneumoniae* and *C. jejuni*. The existence of nitrogen fixation genes in these host-dependent prokaryotes would be difficult to explain: it is unlikely that such an organism has the ability to perform energetically expensive atmospheric nitrogen fixation in the presence of already fixed nitrogen, as in the host environment. Thus, we conclude that the description assigned to these proteins based on the weak similarity to the nitrogen fixation genes is incorrect.

Based on the SNAP prediction, we can conclude that the gene Ta0420 is involved in carbohydrate utilization, possibly as a regulatory protein, which is in accordance with the weak similarity and colinearity data available for this gene.

**Table 2.** Percentage of true positives for individual genomes and summarized for all genomes

| Genome | Percentage of true positives | | Number of genes for which a prediction was made |
|---|---|---|---|
| | Best cycle | All cycles | |
| *A. pernix* | 78.8 | 78.8 | 33 |
| *C. jejuni* | 89.5 | 91.2 | 57 |
| *C. pneumoniae* | 84.2 | 89.5 | 19 |
| *E. coli* | 72.0 | 75.2 | 125 |
| *M. pneumoniae* | 54.5 | 63.6 | 11 |
| *M. thermoautotrophicum* | 76.3 | 76.3 | 38 |
| *M. tuberculosis* | 85.5 | 85.5 | 83 |
| *P. abyssi* | 66.7 | 76.2 | 63 |
| *Synechocystis* sp. | 90.3 | 90.3 | 62 |
| *T. maritima* | 79.6 | 79.6 | 49 |
| *T. pallidum* | 63.6 | 63.6 | 11 |
| *T. acidophilum* | 69.0 | 69.0 | 58 |
| All genomes | 77.8 | 79.8 | 609 |

## Conclusions and Outlook

SNAP is a generalization of the algorithm described by Overbeek *et al*.[11,12] Our method does not rely on the conservation of gene order in the form of colinear gene clusters and detects genes that are functionally coupled through a chain of alternating S and N-relationships. The algorithm takes a protein sequence and a set of annotated completely sequenced genomes as input and returns a number of SN-cycles with all vertices being potentially linked to the query sequence.

The main finding that we report here is the wide occurrence of SN-cycles and their strong non-randomness as compared with genomes in which gene order was artificially shuffled. The fact that SN-cycles actually reflect the conservation of gene order makes them a useful instrument for defining functional relationships among genes, studying genome plasticity, and reconstructing evolutionary events. While the biological background of the SN-cycles remains unclear at this point, we assume that they reflect functional coupling between closely co-regulated genes in prokaryotic genomes and, more generally, the conservation of functional and regulatory contexts in genomes.[18]

Further, we sought to quantify the ability of SNAP to predict broad gene function. Using assignments of genes to KEGG metabolic maps and the genome annotation available through the PEDANT database, we have demonstrated the tendency of SN-cycles to reveal the proximity of functionally coupled genes. In doing so, our consideration was necessarily limited to the genes to which EC numbers could be assigned. Moreover, the metabolic pathway and functional category assignments that served as the basis for calculating the $K_p$ and $K_f$ coefficients were produced automatically based on sequence similarity searches and are prone to errors. Thus, while the anecdotal evidence of functional coupling detection by SNAP presented throughout this work appears to be quite convincing, objective assessment of SNAP performance is very difficult and is currently limited to recovering rough pathway information for some of the genes involved. Moreover, using this approach we are capable of finding putative true positive predictions, but cannot make any conclusions about negative predictions, i.e. cases when no prediction could be made. In any event, it is clear that the reliability of functional inferences made with SNAP will depend critically on the quality of the whole body of genome annotation available.

Significantly better performance of SNAP in terms of the pathway coefficient $K_p$ as compared with the functional category coefficient $K_f$ is not unexpected and is compatible with the main bulk of facts available on the functional composition of gene clusters. Bacterial operons tend to encode members of distinct protein families required for subsequent steps in a biochemical or regulatory pathway. There is also sufficient evidence that the conservation of spatial proximity is especially pronounced between the physically interacting genes.[8,13] We have thus confirmed that the concept of functionally coupled or functionally related genes used in context-based prediction methods actually means functionally interacting or jointly acting genes.

We do not claim to provide the algorithmically most optimal approach to exploring SN-relationships in genomes. The filtering criterion for SN-graphs that we used, namely the requirement for SN-paths to be closed, is essentially equivalent to the requirement of two alternative SN-paths between two functionally coupled genes to be present. A more strict criterion would require that more than two alternative paths between two genes exist. We plan to test the performance of SNAP with the number of gene neighbors in each direction considered $c > 2$ (see Materials and Methods). Increasing $c$ may allow the detection of long-range patterns in gene order.

The main factor limiting the potential of any approach exploiting the conservation of gene order is the massive disruption of gene clusters in distantly related species and the resulting reduction of the number of significant N-relationships available. Another obvious limitation is the possibility of

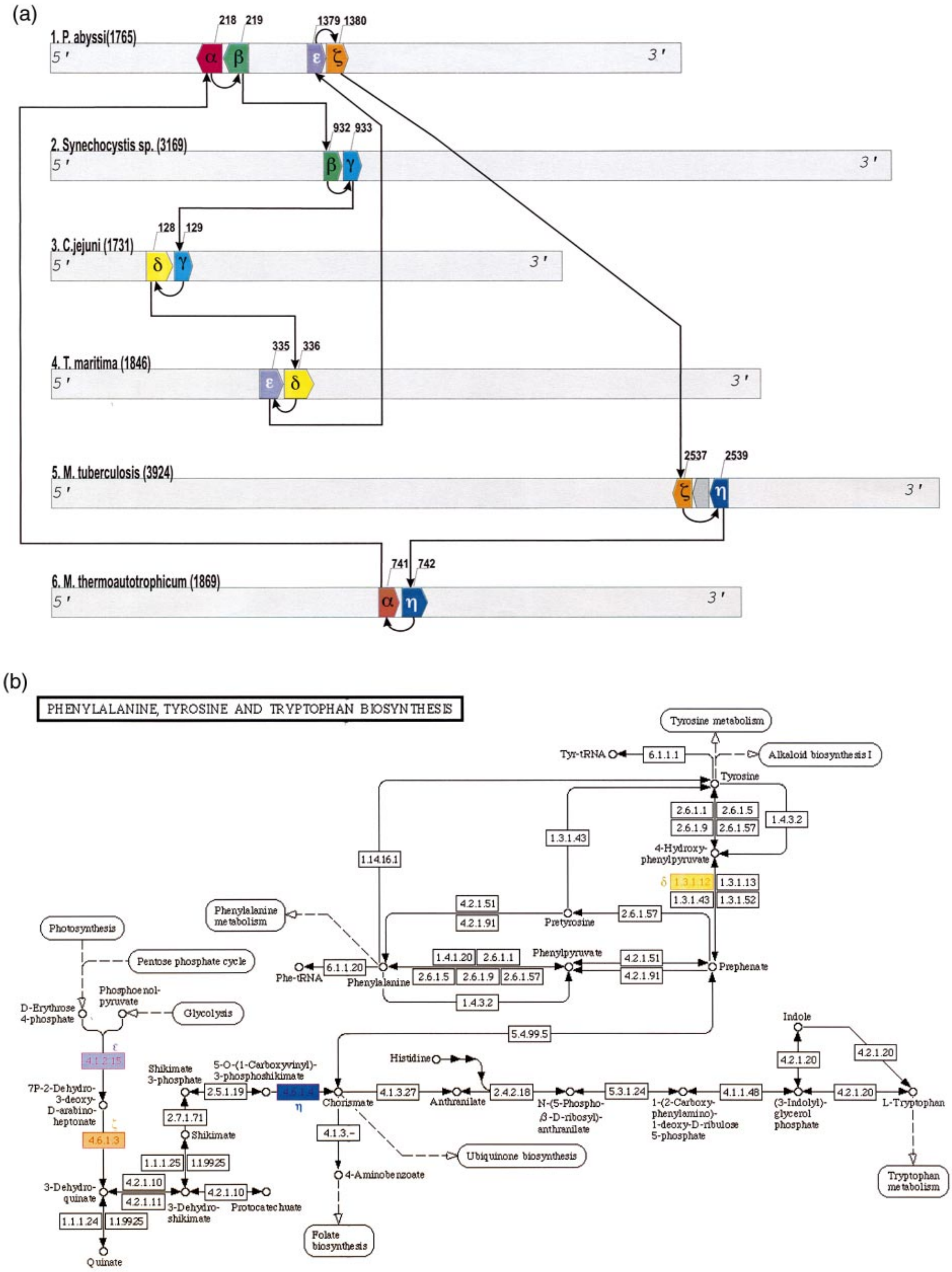

**Figure 6.** SNAP analysis of the hypothetical protein Ta0740 from *T. acidophilum*. (a) SN-cycle associated with Ta0740 (denoted α). Six other protein types found are: β, phosphoribosylaminoimidazolesuccinocarboxyamide synthase (EC 6.3.2.6); γ, chloroplast import-associated channel IAP75; δ, prephenate dehydrogenase (EC 1.3.1.12); ε, 2-dehydro-3-deoxyphosphoheptonate aldolase (EC 4.1.2.15); ζ, 3-dehydroquinate synthase (EC 4.6.1.3); η, chorismate synthase (EC 4.6.1.4). (b) Phenylalanine, tyrosine, and tryptophan biosynthesis pathway as presented in the KEGG database (map 00400). Enzymes δ, ε, ζ, and η are highlighted in colors corresponding to those in (a).

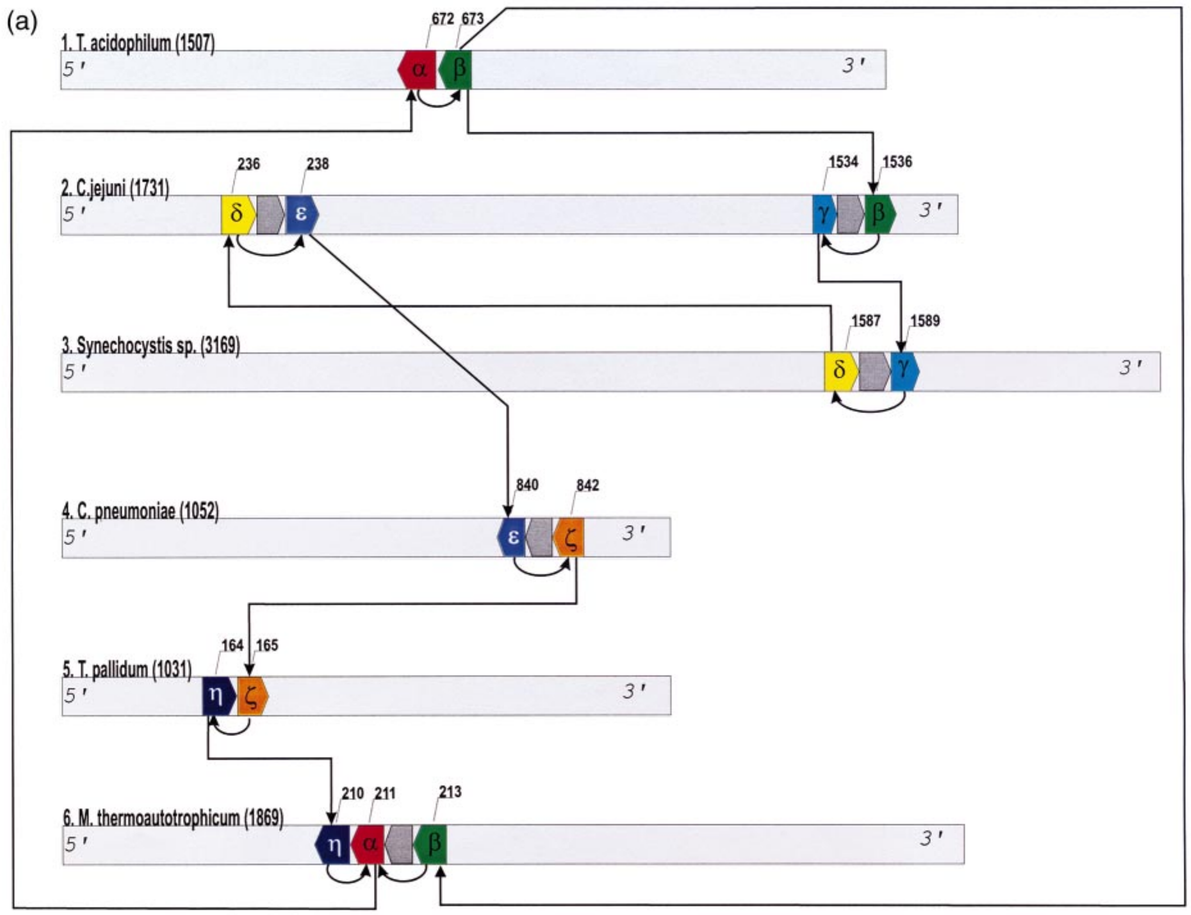

**Figure 7** (*legend shown on page 297*)

non-orthologous gene displacement,[19] leading to termination of SN-cycles due to the absence of their constituent S-relationships. The results of the functional coupling prediction are also dependent on our ability to differentiate orthologs of a certain gene in other genomes from paralogous genes. However, even if a homologous protein with a similar function is recruited instead of the true functional ortholog, the SN-graph may still be closed and the corresponding prediction of significant value.

An important recent advance is the establishing of functional association between spatially separated genes that in other organisms are fused to form a composite protein.[3,4] Gene fusion events have been shown to be reliable indicators of protein interaction, but the number of such events is rather limited (e.g. 64 cases involving 2.8 % of proteins in *E. coli*, *Haemophilus influenzae*, and *Methanoccocus jannaschii*, as reported by Enright *et al.*[4]). It will be easy to adapt SNAP to take into account gene fusion events by redefining N-relationships as those between separate spatially proximate genes, and those between distinct, non-overlapping sequence domains of the same protein as outlined by the structure of BLAST local alignments. SNAP can also be combined with statistical operon prediction methods[20] based on recognition of regulatory DNA signals.

The role and the frequency of occurrence of gene clusters in eukaryotes is completely open. While operons seem not to be generally present in higher organisms, they do play a significant role in some of them. In the *Caenorhabditis elegans* genome, for example, up to 25 % of the genes are organized in polycistronic transcription units.[21] A sizeable number of functionally interacting eukaryotic genes are involved in synexpression groups.[23] What part of these genes are physically associated on the chromosome remains unclear. We intend to study the applicability of our method to the completely sequenced eukaryotic genomes that are currently available.

Based on our tests with the *Thermoplasma acidophilum* genome, we estimate that SNAP will prove instrumental in mapping functional links for a significant fraction (up to 30 %) of presently uncharacterized genes in bacterial genomes. We plan to launch an effort to re-annotate all completely sequenced genomes available to date. Systematic work directed at the detection of functionally interacting genes will have implications for medical

(b)

| Gene | Genome | PEDANT ID | Description line | Automatically assigned functional categories | |
|---|---|---|---|---|---|
| | | | | Number | description |
| α | T. acidophilum | Ta0420 | conserved hypothetical membrane protein | 01.05.04 | regulation of carbohydrate utilization |
| | | | | 02.99 | other energy generation activities |
| | | | | 01.05.01 | carbohydrate utilization |
| α | M. thermoautotrophicum | gi_2621261 | FUN34 related protein | 01.05.04 | regulation of carbohydrate utilization |
| | | | | 02.99 | other energy generation activities |
| | | | | 01.05.01 | carbohydrate utilization |
| β | T. acidophilum | Ta0421 | probable acetyl-coenzyme-A synthetase | 02.99 | other energy generation activities |
| | | | | 01.05.01 | carbohydrate utilization |
| | | | | 30.16 | mitochondrial organization |
| | | | | 09.01 | biogenesis of cell wall |
| β | C. jejuni | cj1537c | acetyl-CoA synthetase | 02.99 | other energy generation activities |
| | | | | 01.05.01 | carbohydrate utilization |
| | | | | 30.16 | mitochondrial organization |
| | | | | 09.01 | biogenesis of cell wall |
| γ | C. jejuni | cj1535c | glucose-6-phosphate isomerase | 30.03 | organization of cytoplasm |
| | | | | 02.01 | glycolysis |
| | | | | 02.04 | gluconeogenesis |
| | | | | 01.05.01 | carbohydrate utilization |
| γ | Synechocystis sp. | gi_1653253 | glucose-6-phosphate isomerase | 30.03 | organization of cytoplasm |
| | | | | 02.01 | glycolysis |
| | | | | 02.04 | gluconeogenesis |
| | | | | 01.05.01 | carbohydrate utilization |
| δ | Synechocystis sp. | gi_1653251 | carbonic anhydrase | 08.16 | extracellular transpor |
| δ | C. jejuni | cj0237 | carbonic anhydrase | 08.16 | extracellular transpor |
| ε | C. jejuni | cj0239c | nifU-like protein | 01.02.01 | nitrogen and sulphur utilization |
| ε | C. pneumoniae | gi_4377178 | NifU-related protein | 01.02.01 | nitrogen and sulphur utilization |
| ζ | C. pneumoniae | gi_3322436 | phosphoglycerate mutase | 01.05.01 | carbohydrate utilization |
| | | | | 02.01 | glycolysis |
| | | | | 30.03 | organization of cytoplasm |
| ζ | T. pallidum | gi_3322436 | phosphoglycerate mutase | 01.05.01 | carbohydrate utilization |
| | | | | 02.01 | glycolysis |
| | | | | 30.03 | organization of cytoplasm |
| η | T. pallidum | gi_3322435 | cation-activated repressor protein | - | - |
| η | M. thermoautotrophicum | gi_2621260 | iron dependent repressor | - | - |

**Figure 7.** SNAP analysis of the hypothetical protein Ta0420 (α) from *T. acidophilum*. (a) SN-cycle associated with Ta0740 (denoted α). (b) Functional categories assigned by PEDANT.

and environmental research, since many genes responsible for antibiotic resistance, pathogenesis, and biodegradation are transferred horizontally between different species in clusters[23] and consequently represent good targets for SNAP. A WWW server allowing the users to perform a gene func-

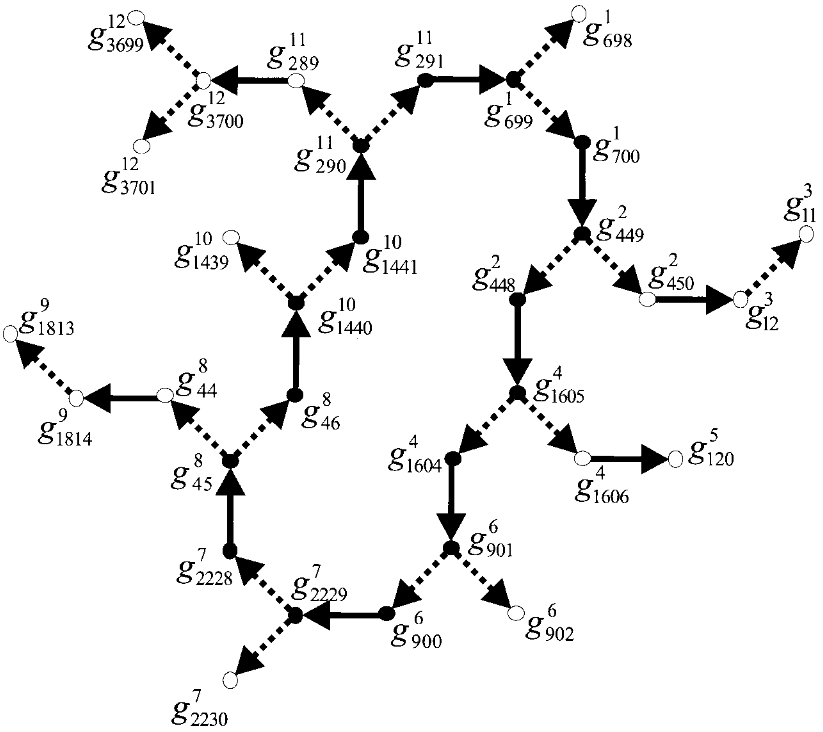

**Figure 8.** A hypothetical chain of SN-relationships. A part of a hypothetical SN-graph involving an SN-cycle. Genes participating and not participating in the SN-cycle are shown as filled and open circles, respectively, and are denoted as $g^i_k$ where the superscript stands for the genome number and the subscript for the sequential gene number on the chromosome. Continuous and broken arrows depict similarity and neighborhood-relationships, respectively. The number of gene neighbors considered on each side $c = 1$.

tion prediction using our method and the underlying PEDANT genome database is now under development.

## Materials and Methods

### Description of the algorithm

We consider $N$ bacterial genomes $G_i$ ($i = 1,N$), each containing $M^i$ genes, $g^i_k(k = 1, M^i)$, where $k$ is the sequential number of the gene on the chromosome. Two genes, $g^i_k$ and $g^i_{k+1}$, from the same genome $i$ are N-related if they fulfil the following conditions: (i) the distance between the stop codon of $g^i_k$ and the start codon of $g^i_{k+1}$ is smaller then a certain threshold value $d$ (typically 500 base-pairs.); and (ii) both $g^i_k$ and $g^i_{k+1}$ have the same orientation (i.e. they are situated on the same strand; as demonstrated by Overbeek *et al.*,[12] co-occurence of functionally coupled genes on opposite strands is a very rare event). We take into account spatial association between genes that are, at most, $c$ genes away from each other. Therefore, a genome $i$ can be represented as an unordered set of up to $M^i - 2c$ gene words, $W^i_q(q = c + 1,M^i - c)$, each word being an ordered list of up to $2c + 1$ genes:

$$W^i_{c+1}=\langle g^i_1\cdots g^i_{2c+1}\rangle,\ W^i_{c+2}=\langle g^i_2\cdots g^i_{2c+2}\rangle,\ W^i_{c+3}=\langle g^i_3\cdots g^i_{2c+3}\rangle$$

etc. In other words, each gene word $W^i_q$ contains the gene $g^i_q$, its $c$ neighbors on the left, and its $c$ neighbors on the right. A genome will contain exactly $M^i - 2c$ gene words only if all genes are on the same strand and are separated by no more than $d$ bases. Since this is never the case, the actual number of gene words in a genome will be smaller. For the same reason, many of the gene words will contain less than $2c + 1$ genes. The minimal number of genes in a gene word is two, since otherwise no N-relationship in the word can exist. Throughout this work we used $c = 2$ in order to make our tests computationally feasible.

An all-against-all comparison of the genes $g^i_k$, ($i = 1,N$, $k = 1,M^i$) is conducted using the PSI-BLAST algorithm.[16] An S-relationship between two genes $g^i_k$ and $g^j_l$, residing on the genomes $G_i$ and $G_j$, respectively, exists if the BLAST $E$-value $E(g^i_k, g^j_l) < e$, and the coverage of the BLAST alignment, defined as the fraction of amino acid residues of the shorter compared protein covered by the alignment, $C(g^i_k, g^j_l) > a$, where $e$ and $a$ are parameters of the analysis. As an additional restriction, we may require the BLAST match to be reciprocal, such that $E(g^i_k, g^j_l) < e$, $E(g^j_l, g^i_k) < e$ and there are no $x = 1,M^j$, $x \neq k$ and $y = 1,M^i$, $y \neq l$ such that $E(g^i_x, g^j_k) < E(g^j_l, g^i_k)$ and $E(g^i_y, g^j_l) < E(g^i_k, g^j_l)$. The matrix of all-against-all BLAST matches is made symmetrical by selecting for each pair of proteins the best $E$-value and the best value of coverage $C$, such that $E(g^i_k, g^j_l) = E(g^j_l, g^i_k) = \min(E(g^i_k, g^j_l), E(g^j_l, g^i_k))$ and $C(g^i_k, g^j_l) = C(g^j_l, g^i_k) = \max(C(g^i_k, g^j_l), C(g^j_l, g^i_k))$.

We can now represent the chain of SN-relationships originating from an arbitrary gene $g^i_k$ as an SN-graph involving S and N-relationships in an alternating fashion, starting either with an S-relationship or an N-relationship in which $g^i_k$ is involved. An example of such a graph is shown in Figure 8. It is easy to see that the SN-graph joins gene words that have at least one pair of S-related genes.

In our implementation, an SN-graph is traversed using the depth-first algorithm and all closed SN-paths, or SN-cycles, are identified. In Figure 8, an SN-cycle involves 16 genes shown as filled circles, corresponding to the eight related gene words. A special case of an SN-cycle is constituted by colinear gene clusters in which the order of genes is partially or fully conserved across several genomes. Such SN-cycles involve words with more than one pair of S-related genes (Figure 9).

With an increasing number of genomes the number of nodes in the SN-graph grows very quickly so that find-

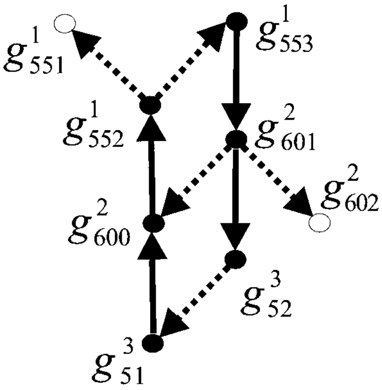

**Figure 9.** A hypothetical SN-graph which involves a conserved pair of genes in three genomes: genes 552 and 553 in genome 1, genes 600 and 601 in genome 2, and genes 51 and 52 in genome 3. In this case, the SN-cycle is equivalent to a colinear gene cluster of the type described by Overbeek *et al.*[12] The notation as in Figure 8.

ing all paths becomes computationally prohibitive. To demonstrate the feasibility of our approach, without losing the generality, we set an upper limit on the path length at a certain value, typically 14 nodes.

### Measuring the performance of the method

All genes belonging to an SN-cycle are regarded as functionally coupled. In order to test the validity of this assertion, we need to measure the performance of the algorithm on a large number of documented cases of "true" functional relatedness. Two different approaches for defining the standard of truth for our calculations have been explored.

#### *Analysis of reference metabolic pathways*

The entire KEGG/PATHWAY database†,[17] was processed with a sophisticated perl script to extract the pathway graph in a form suitable for subsequent computer analysis. Information about links between biological objects cannot be gleaned easily from the KEGG image files representing the pathways. We obtained this information indirectly by comparing the list of all biochemical reactions present in the database with another list that specifies both the EC number of a given enzyme and the compounds it interacts with. Since the names of the compounds in the first and the second list are often inconsistent, we used a sub-string comparison technique to establish correspondence between them. Further, unspecific widely applicable metabolites, such as water, alcohol, $CO_2$ etc. were not considered.

The pathway graph is constituted by vertices and edges corresponding to enzymes and substrates, respectively. Given a set of enzymes represented by their EC numbers $E = (E_1, E_2, \ldots, E_n)$, where $n$ is the number of enzymes in the set, our goal is to find a measure, $0 \leqslant K_p \leqslant 1$, to describe their "concentration" on the pathway graph. We call this measure the "pathway coefficient". The ideal case of $K_p = 1$ corresponds to an SN-cycle joining enzymes that form a compact pathway subgraph such that (i) no other nodes except for $(E_1, E_2, \ldots, E_n)$ exist, and (ii) for any nodes $E_i$ and $E_j$ there exists a path connecting them. The worst case $K_p = 0$ describes an SN-cycle that joins totally unrelated enzymes, i.e. there is no path on the pathway graph connecting any pair of the enzymes found.

The metabolic distance $D_{ij}$ between two enzymes $E_i$ and $E_j$ on the pathway graph is defined as the minimal number of reaction stages (edges) connecting these enzymes (vertices). Given a set of enzymes, we used the following approach to determine the value of the pathway coefficient $K_p$. Single linkage clustering was applied to the metabolic distance matrix $D_{ij}$, $i = 1,n$, $j = 1,n$ in order to find the largest cluster of vertices $C \in E$ subject to the constraint that $D_{ij} < D_t$, where $D_t$ is the threshold metabolic distance. The pathway coefficient can then be computed as:

$$K_p = \lambda_p \frac{m}{n} \qquad (1)$$

where $m$ is the number of elements in $C$, and $\lambda$ is a normalization coefficient defined as:

$$\lambda_p = \frac{m}{\sum_{j=1}^{m} q_j} \qquad (2)$$

where $q_j$ denotes the number of times the EC number corresponding to the $j$th element of $C$ occurred in the entire pathway graph.

#### *Utilization of functional categories*

The degree of functional coupling between the genes involved in SN-cycles was also examined in reference to the MIPS functional role catalogue developed for the yeast genome‡.[24] The catalogue has a hierarchical structure. Each of the 15 main classes (e.g. metabolism, energy etc.) contains three to four subclasses, with the total number of functional categories exceeding 200. Correspondingly, the numeric designator of a functional class can include up to four numbers. For example, the yeast gene product YGL237c is attributed to the functional category 04.05.01.04, where the numbers, from left to right, mean transcription, mRNA transcription, mRNA synthesis, and transcriptional control. Nearly 4000 yeast genes could be ascribed to at least one functional category based on careful manual analysis of extrinsic evidence (similarity to known proteins, presence of indicative sequence patterns) as well as experimental data from the literature. In this work, the MIPS classification was used for automatic assignment of functional categories to gene products from completely sequenced genomes based on significant homology to one or many functionally characterized yeast genes.

The functional category coefficient for a group of genes with at least one functional category assigned $F = (F_1, F_2, \ldots, F_n)$ was computed as:

$$K_f = \lambda_f \frac{m}{n} \qquad (3)$$

† Available online at http://www.genome.ad.jp/kegg/kegg2.html, downloaded from ftp://kegg.genome.ad.jp

‡ Available online at http://mips.gsf.de/proj/yeast/catalogues/funcat/index.html

where $n$ is the number of genes in the group, $m$ is the maximal number of times a functional category $f$ occurred in $F$, and $\lambda_f$ is a normalization coefficient:

$$\lambda_f = 1 - P(m,f) \tag{4}$$

In the latter equation $P(m,f)$ denotes the binomial probability of the functional category $f$ to occur $m$ times in the group of genes of size $n$:

$$P(m,f) = \frac{n!}{(n-m)!} p^m (1-p)^{n-m} \tag{5}$$

where $p$ is the general frequency of occurrence of a functional category $f$.

### Implementation and data sources

The main vehicle for the present study was the PEDANT genome analysis system.[25,26] The PEDANT database† contains exhaustive functional and structural annotation of all completely sequenced genomes. In particular, gene products are automatically assigned to yeast functional categories[24] and enzyme classes[17] based on similarity searches. Out of 35 finished genomic sequences available at the time of writing, we selected 12 genomes from sufficiently distant species, as assessed visually based on a maximum likelihood phylogenetic tree derived from the small-subunit rRNA sequences using the PHYLIP package (data not shown).[27] Namely, these genomes are: *Aeropyrum pernix, C. jejuni, C. pneumoniae, E. coli, M. pneumoniae, M. thermoautotrophicum, Mycobacterium tuberculosis, Pyrococcus abyssi, T. acidophilum, T. maritima, T. pallidum, Synechocystis* sp.‡. Throughout this text, gene IDs as available through the PEDANT database are utilized.

A perl program was written to extract gene positional information and various other attributes from the PEDANT MySQL relational tables, build the SN-graphs, detect SN-cycles, and study the features of the genes predicted to be functionally related.

## Acknowledgments

We thank Matthias Fellenberg, Jean Hani, and Stephen Rudd for careful reading of the manuscript and many useful comments. This work was supported by Deutsche Forschungsgemeinschaft (DFG).

## References

1. Andrade, M. A. & Bork, P. (2000). Automated extraction of information in molecular biology. *FEBS Letters*, **476**, 12-17.
2. des Jardins, M., Karp, P. D., Krummenacker, M., Lee, T. J. & OuzounisC., A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Intell. Syst. Mol. Biol.* **5**, 92-99.
3. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83-86.
4. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86-90.
5. Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V. & Riley, M.*et al*. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1462.
6. Mushegian, A. R. & Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends Genet.* **12**, 289-290.
7. Watanabe, H., Mori, H., Itoh, T. & Gojobori, T. (1997). Genome plasticity as a paradigm of eubacteria evolution. *J. Mol. Evol.* **44**, S57-S64.
8. Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16**, 332-346.
9. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66-73.
10. Bansal, A. K. (1999). An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics*, **15**, 900-908.
11. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1998). Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* **1**, 0009.
12. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896-2901.
13. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci.* **23**, 324-328.
14. Lawrence, J. G. & Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843-1860.
15. Huynen, M., Snel, B., Lathe, W., III & Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204-1210.
16. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
17. Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27-30.
18. Lathe, W. C., Snel, B. & Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* **25**, 474-479.
19. Koonin, E. V., Mushegian, A. R. & Bork, P. (1996). Non-orthologous gene displacement. *Trends Genet.* **12**, 334-336.
20. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. (2000). A probabilistic learning approach to whole-genome operon prediction. *Intell. Syst. Mol. Biol.* **8**, 116-127.
21. Blumenthal, T. & Spieth, J. (1996). Gene structure and organization in *Caenorhabditis elegans*. *Curr. Opin. Genet. Dev.* **6**, 692-698.
22. Niehrs, C. & Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature*, **402**, 483-487.

† Available online at http://pedant.gsf.de

‡ URLS for the respective sequencing centres are available at http://pedant.gsf.de/credits.html

23. De La Cruz, I. & Davies, I. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8**, 128-133.

24. Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J. *et al*. (1997). Overview of the yeast genome. *Nature,* **387**, 7-65.

25. Frishman, D. & Mewes, H. W. (1997). PEDANTic genome analysis. *Trends Genet.* **13**, 415-416.

26. Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. & Mewes, H.-W. (2000). Functional and structural genomics using PEDANT. *Bioinformatics.*

27. Felsenstein, J. (1989). PHYLIP - phylogeny inference package. *Cladistics,* **5**, 164-166.

28. Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W. *et al*. (2000). The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature,* **407**, 508-513.

***Edited by J. Thornton***

*(Received 25 January 2001; received in revised form 6 April 2001; accepted 12 April 2001)*