

Discovery of Frequent Gene Patterns in Microbial Genomes

Tom H. Wetjen

November 12, 2002

TZI-Bericht 27, November 2002

*Center for Computing Technologies (TZI), University of Bremen
P.O. Box 330 440, D-28834 Bremen
T. ++49 421 218 7838, twetjen@tzi.de*

1 Introduction

The accumulation of complete sequences of many microbial genomes (bacterial and archaeal) opens new opportunities for a more comprehensive comparison of microorganisms. The comparison of whole genome sequences allows the identification of coding and regulatory regions, the function and the evolution of gene families, the rate of gene conservation, variations of gene functionality in various organisms, and mechanisms of evolution [1, 2]. Particularly, the interspecies comparison of gene order got into focus in the last years. Information received from gene cluster analysis will help identifying functionally related genes and thus, support the understanding of metabolic pathways. Furthermore, the rate of gene order conservation improves the development of evolutionary hypotheses about events like horizontal gene transfer and specification [3].

The discovery of conserved gene clusters is a non-trivial task, since microbial genomes can have up to nearly 10,000 genes [4]. Furthermore, the architecture of microbial genomes, even within the same species, can be highly variable and the order of genes and their coding is less conserved as could be expected. Moreover, evolutionary processes causing this fluidity are not fully understood yet [5].

There exist several methods which have been proposed for comparing gene orders in pairs of genomes, multiple genomes, and for detecting local gene order conservation [6, 7, 8, 9, 3, 10]. These methods differ with respect to the amount of gene insertions/deletions and local rearrangements allowed. The application of these and other methods have produced a wealth of functional and evolutionary information, which was interpreted in the more general framework of genome

context analysis. However, all methods available at present lack the ability to discover qualitative relations about gene order conservation which in turn can be used to improve the genome context analysis. For example, the alignment of microbial genomes can be supported by spatial knowledge of genomic structures. A qualitative representation is required since quantitative modeling of genome structures (concrete positions, lengths, and distances) would disregard genome rearrangements and any kind of mutations. Thus, an approach is required which discovers gene order conservation and, furthermore, qualitative spatial relations of gene order.

Here we will describe an approach to discover frequent gene patterns in microbial genomes which, in addition to the identification of conserved gene clusters, is capable of extracting spatial relations from the data. The approach is based upon the combination of a system of interval relations known from AI research on temporal and spatial reasoning [11], and upon the application of association rules [12] on instances modeled by these interval relations.

2 Material and Methods

2.1 Spatial-Rational Modeling of Gene Order

Microbial genome sequences are assumed to be linear as in data base entries instead of circular as for most microorganisms in *vivo*. Any gene is modeled as an interval in the form of $s_i < e_i$, with s_i and e_i as defined starting- and end-points in the genome sequence. For simplification the genes are continuously numbered in the 5' \rightarrow 3'-direction from 1 to n in their order of succession. A genome sequence containing n genes is of the form $\sigma : (g_1(s_1, e_1), \dots, g_n(s_n, e_n) | s_i < s_j)$.

To qualitatively model spatial relations between genes we use Allen's interval logic [11]. There exists a set of thirteen basic relationships denoted by I which can hold between two intervals, namely: before (b), after (bi), meets (m), met-by (mi), overlaps (o), overlapped-by (oi), starts (s), started-by (si), finishes (f), finished-by (fi), during (d), contains (di), and equals (eq) (Figure 1).

Given n genes of a genome sequence σ , we can capture their relative positions to each other by an $n \times n$ matrix N where any cell $N(i, j)$ describes the relationship $r \in I$ between the genes i and j . Thus, a genome instance is a tuple $\gamma : (\sigma, N)$.

2.2 Discovery of Frequent Gene Patterns

A gene pattern π of size k is defined by a triplet $\pi : (p, R, \tau)$, where $p : \{g_1, \dots, g_k\}$ is the set of genes included, $R \in I^{n \times n}$ denotes the relationships between any $g_i(s_i, e_i)$ and $g_j(s_j, e_j)$, and $\tau \subseteq \Gamma$, where Γ is the set of all genomes under evaluation. We call such a pattern by definition a k -pattern. The frequency f_π of a gene pattern π is its number of distinct occurrences in the set of genome


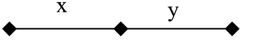
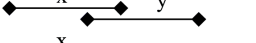
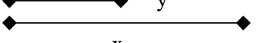
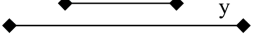
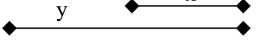

Relation	Inverse	Meaning
x before y	y after x	
x meets y	y met-by x	
x overlaps y	y overlaped-by x	
x starts y	y started-by x	
x during y	y contains x	
x finishes y	y finished-by x	
x equals y		

Figure 1: The 13 basic relations which can hold between two intervals.

instances under evaluation. In order to be considered relevant, a gene pattern has to contain a minimum number of genes, i.e. k has to exceed a given threshold. Furthermore, only those patterns which can be observed in a neighborhood with a specified number w of genes are generated. This in turn limits the number k of genes in any π to a maximum of w .

To find all patterns of a set of genome instances Γ whose frequency lies above a given threshold, first all frequent 1-patterns ($k = 1$) are constructed. Next we try to extend any 1-pattern to a 2-pattern by iterating over its associated set τ of genomes and searching for frequent genes within w . The fact that the frequency of a pattern is less than or equal to the frequency of any of its subpatterns guarantees that no frequent pattern will be missed [13]

$$\forall \pi, \omega : \omega \subset \pi \Rightarrow f(\omega) \geq f(\pi) \quad (1)$$

With the increase of k , the number of potential patterns grows exponentially. Thus, constructing all k -patterns as described above can be very time-consuming for large k 's. Therefore, pruning techniques introduced by Höppner [14] are used to keep the increase in the number of candidate patterns moderate. Every subpattern of a $(k+1)$ -pattern candidate is frequent after (1). Thus, any two frequent k -patterns π_i and π_j can be joined to a $(k+1)$ -pattern π_x , if they occur in the same genome instances and share a common subpattern. Let us denote the remaining genes beside the common subpattern in π_i and π_j with p and q , respectively. In order to build the relation matrix of π_x the relations for the common subpattern denoted by A can be taken from π_i or π_j . Furthermore, the relations between p and q and the first $(k-1)$ gene can also be taken from π_i and π_j . Thus, the only degree of freedom within π_x is the relation r between p and q . Figure 2 illustrates how to build the $(k+1)$ -pattern matrix π_x out of π_i and π_j .

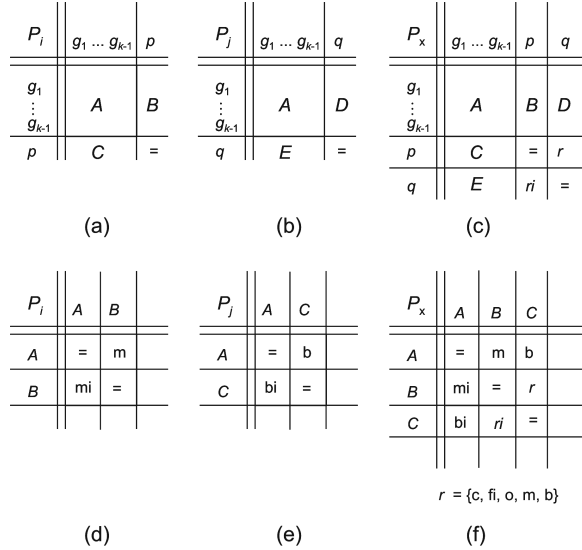


Figure 2: General procedure to generate a $(k+1)$ -pattern π_x (c) out of two k -patterns π_i (a) and π_j (b). An example of the procedure is given in (d) (e) (f), where the common subpattern of (d) and (e) is A.

The freedom in choosing r yields 13 different patterns which may become the candidate $(k+1)$ -pattern because there are 13 interval relations. Since the search for an extension is directed into the $5' \rightarrow 3'$ -direction, we can reduce the possible values of r to a maximal number of 7 by ignoring the inverse relations without any loss of generality. Before checking all genome instances of the k -patterns for one of the relations between p and q , another pruning technique based on the law of transitivity can be applied. In Allen [11] a transitivity table for any relations between 3 intervals is given where the relation r_{ab} between a and b , and r_{bc} between b and c may constrain the possible relations r_{ac} between a and c . For example, the 2-patterns "A meets B" and "A before C" share the $(k-1)$ -pattern "A". The missing relation r between B and C is $\{c, fi, o, m, b\}$, obtained by the transitivity table. Only those relations which do not contradict the result obtained by the law of transitivity can be included in the $(k+1)$ -pattern π_x . All k -patterns sharing a common subpattern and occurring in the same instances can recursively be joined to $(k+1)$ -patterns until either all k -patterns having a $k = w$, or there exists no further two k -patterns which can be joined. Patterns generated are stored according to their k in decreasing order of k . Furthermore, all k -patterns are sorted in decreasing order of their frequency f .

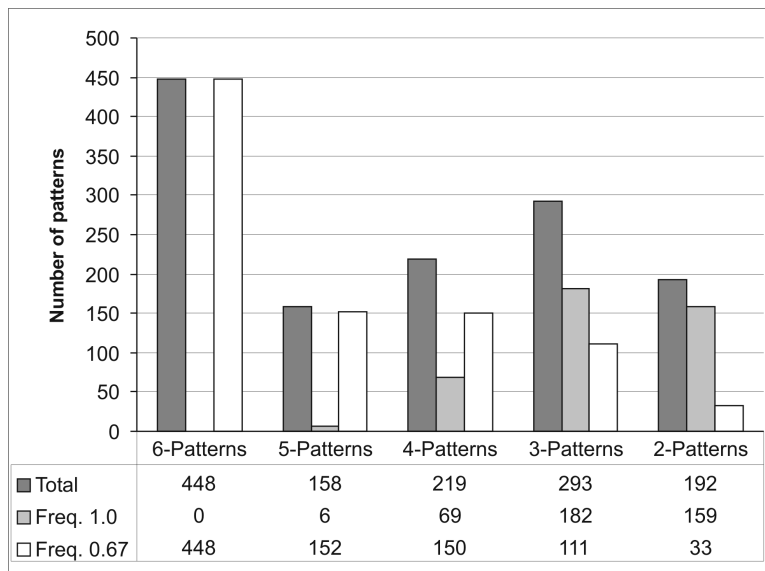


Figure 3: Number of discovered gene patterns between Mge, Mpn and Uur.

2.3 Producing a Non-Redundant Set of Patterns

Since patterns are generated by extending any k -pattern in the $5' \rightarrow 3'$ -direction the result-set of the frequent patterns may contain some redundancy. This redundancy occurs if there exist a k -pattern π_i starting with a gene g_i , with $s_i < \dots < s_k$, and another k -pattern π_j starting with the gene g_{i+1} , with $s_{i+1} < \dots < s_k$, i.e. the pattern π_j is a subpattern of π_i . In order to produce a non-redundant set of patterns, for any k -pattern it is tested whether there exists a $k+n$ -pattern, with $n = w - k$, of which the k -pattern is a subpattern. If so, this k -patterns is removed from the result set.

3 Evaluation

3.1 Sequence Data

For an evaluation, the approach was applied to orthologous genes of 3 complete prokaryotic genomes which were extracted from the database of Clusters of Orthologous Genes Groups of Proteins(COGs) at the Genome division of the NCBI (<http://www.ncbi.nlm.nih.gov/COG/>) [15, 16]. The analyzed genomes were of the following species: *Mycoplasma genitalium* (Mge), *Mycoplasma pneumoniae* (Mpn), and *Ureaplasma urealyticum* (Uur), all belonging to the family of *Mycoplasmataceae*.

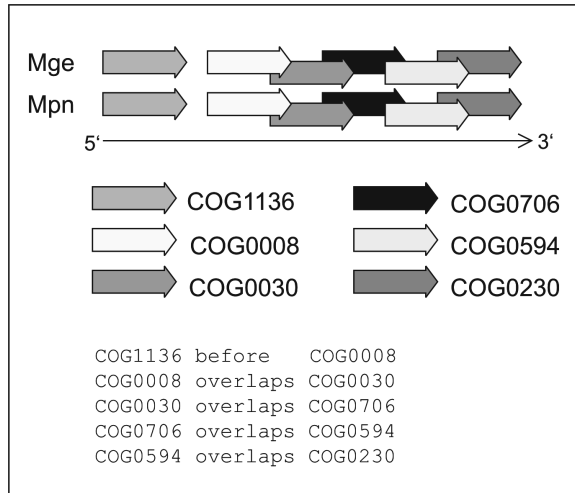


Figure 4: Example of a 6-pattern between Mge and Mpn. The pattern is situated on the minus strand. The length and distances of the COGs are not in scale!

3.2 Results

Adjusting the parameters to the following: $k = 2$ (at least two genes in any pattern), $w = 6$ (max. 6 genes in any pattern), and $f = 2$ (patterns occur in at least two genomes), altogether, 1310 conserved clusters of orthologous genes were detected. Only 416 of these clusters could be observed in all 3 genomes (frequency = 1.0), which in turn were distributed to 159 2-patterns, 182 3-patterns, 69 4-patterns, and 6 5-patterns (Figure 3). All 6-patterns observed showed a frequency of 0.67, i.e., were detected in only two genomes. Moreover, only three of the 6-patterns were detected between Mge and Uur, whereas all other 442 were observed between Mge and Mpn. This is probably due to the fact that Mge and Mpn are phylogenetic closer related to each other than is Uur to either Mge or Mpn. Figure 4 gives an example of a 6-pattern identified between Mge and Mpn.

4 Conclusions

We have proposed a technique for the discovery of gene patterns from a set of complete microbial genomes. The example in section 3 has shown that the proposed method is capable of discovering large, complex, and conserved clusters of genes. In addition, the approach will find qualitative relations between those genes. In order to model gene order, the approach uses intuitive interval relations which allows for an easy verification of the knowledge by an expert. Furthermore, the knowledge gained through the approach can be introduced without further remodeling within knowledge-based systems since they allow for a formal reasoning.

References

- [1] J. Overbeek, C. R. Woese, and R. Overbeek. The winds of evolutionary change: Breathing new life into microbiology. *Journal of Bacteriology*, 176(1):1–6, 1994.
- [2] Arvind K. Bansal, Peer Bork, and Peter J. Stuckey. Automated pair-wise comparisons of microbial genomes. *Math. Modelling and Sci. Computing*, 9(1):1–23, 1998.
- [3] Ross Overbeek, Michael Fonstein, Mark D’Souza, Natalia Maltsev, and Gordon D. Pusch. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci.*, 96:2896–2901, 1999.
- [4] Sherwood Casjens. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.*, 32:339–377, 1998.
- [5] Siv G. E. Andersson and Kimmo Eriksson. Dynamics of gene order structures and genomic architectures. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, volume 1 of *Computational Biology Series*, pages 267–280. Kluwer Academic Publisher, London, 2000.
- [6] Igor B. Rogozin, Kira S. Makarova, Janos Murvai, Eva Czabarka, Yuri I. Wolf, Roman L. Tatusov, Laszlo A. Szekely, and Eugene V. Koonin. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Research*, 30(10):2212–2223, 2002.
- [7] Steffen Heber and Jens Stoye. Algorithms for finding gene clusters. In O. Gascuel and B. M. E. Moret, editors, *WABI 2001*, volume 2149, pages 252–263, Aarhus, Denmark, 2001. Springer Verlag.
- [8] Raja Mazumder, Ashok Kolaskar, and Donald Seto. Geneorder: Comparing the order of genes in small genomes. *Bioinformatics*, 17(2):162–166, 2001.
- [9] Yuri I. Wolf, Igor B. Rogozin, Alexey S. Kondrashov, and Eugene V. Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research*, 11:356–372, 2001.
- [10] Maria D. Ermolaeva, Owen White, and Steven L. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Research*, 29(5):1216–1221, 2001.
- [11] James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

- [12] R. Agrawal, T. Imielenski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOND Conference on Management of Data (SIGMOND'98)*, pages 207–216, New York, USA, 1993. ACM Press.
- [13] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. MIT Press, 1996.
- [14] Frank Höppner. Discovery of temporal patterns. In *Proceedings of the 5th European Conference on Principals and Practices of Knowledge Discovery in Databases*, volume 2168 of *Lecture Notes in Artificial Intelligence*, pages 192–203, Freiburg, Germany, 2001. Springer.
- [15] Roman L. Tatusov, Eugene V. Koonin, and David J. Lipman. A genomic perspective on protein families. *Science*, 278(1):631–637, 1997.
- [16] Roman L. Tatusov, Darren A. Natale, Igor V. Garkavtsev, Tatiana A. Tatusov, Uma T. Shankavaram, Bachoti S. Rao, Boris Kiryutin, Michael Y. Galperin, Natalie D. Fedorova, and Eugene V. Koonin. The cog database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1):22–28, 2001.