DIMACS Workshop on Whole Genome Comparison
February 28 - March 2, 2001
DIMACS Center, Rutgers University, Piscataway, NJ, USA

Detection of related genes in procaryotes using syntenic regions (extended abstract)

Nalvo F. de Almeida Jr. *
DCT-UFMS, CP 549, Campo Grande,MS, 79070-900, Brazil.
nalvo@dct.ufms.br

João Carlos Setubal
IC-UNICAMP, CP 6176, Campinas, SP, 13083-970, Brazil.
setubal@ic.unicamp.br

In this work we explore the synteny between prokaryotic genomes. One of our goals is to obtain clues for the function of hypothetical genes; another is to gain a better understanding of changes in gene organization between different species. Our method is based on a careful pairwise comparison of genes in whole genomes.

We first build a binary similarity matrix M based on BLAST e-values of genome G genes against genome H genes and vice versa (we use default BLAST parameter values, but do not use the low-complexity region filter). An entry m_{gh} in M is 1, for gene g in G and gene h in H, if and only if the e-values resulting from having g as query against h and h as query against g are both $\leq 10^{-5}$. We refer to each $m_{gh} \neq 0$ as a *match*. We then look for physically (in each genome) consecutive matches (called *runs*) and for runs that are within k of each other, where k measures the maximum number of intervening genes in one genome, or in the other, or in both. Runs are found by simple scanning of match lists. Note that there are two types of runs: parallel (both sets of genes are in increasing order in their respective genomes) and antiparallel (one is increasing and the other is decreasing). Note also that a run is more general than an operon. In an operon, a series of consecutive genes is transcribed as a unit, and hence must all be on the same strand. Our concept of run allows consecutive genes in one or the other genome to be in different strands. When we find runs of the same type that are within k and such that the flanking genes in each run have consistent strands we join them. This join operation adds the intervening genes in both genomes to the run. We use the term *clusters* to designate joined runs.

We designate a pair of genes (g, h), g in G and g in H, inside a cluster, such that nor g nor h belong to a run, as a *candidate related pair* (CRP). Some of these pairs will be genes that are evolutionary related, but whose similarity cannot be detected by the usual homology methods. Some of the criteria that can be used to build the case that the pair (g, h) is indeed related are:

- Relative size – One would expect that functionally related genes have approximately the same size (the ratio of the smaller to the larger being no less than, say, 80%), but this is by no means an universal rule.
- Consistent strands – The joining of runs impose a constraint on the strands of genes belonging to the cluster. In a parallel run, the strands of each pair (g, h) belonging to the run must match. In an antiparallel run, the strands must be of opposite sign (taking one strand as + and the other as −). We would expect the same consistency to occur in CRPs that are evolutionary related, but again there may exceptions to this.

We have obtained two other kinds of comparative information using our method. The first is related to fusion and fission of genes. A candidate gene fusion event occurs when gene i in G has matches

*Funding from UFMS, FUNDECT, and CNPq/PRONEX[664107/1997-4].

to genes j and $j+1$ in H. Since our definition of match implies reciprocity, genes j and $j+1$ are a candidate gene fission event. We again use the word "candidate" to stress the fact that each of these results must be carefully analyzed to determine whether they correspond to actual biological events.

A second kind of information is a generalization of fusion/fission using the concept of runs. We can detect the following kind of event: given two parallel runs $A = ((g_i, h_j), (g_{i+1}, h_{j+1}), \ldots, (g_k, h_l))$ and $B = ((g_p, h_q), (g_{p+1}, h_{q+1}), \ldots, (g_r, h_s))$, we also have that $p = k+1$ but $l < q+1$. In other words we have a consecutive series of genes in G that is split in two in H. An analogous definition can be given for two antiparallel runs. We term such events as candidate run fusion events (from the point of view of G) and candidate run fission events (from the point of view of H).

We present results of using our method on the publicly available genomes *Xylella fastidiosa* (XF), *Escherichia coli* (EC) and *Pseudomonas aeruginosa* (PA). We also use the current version of the *Xanthomonas axonopodis pv citri* (XC) genome, which is now being finished by a consortium of labs in Brazil. A summary of the results we obtained follows in the tables below.

Table 1: Matches, runs, and clusters (using $k = 2$)

organisms	matches	runs	clusters
XF vs. XC	5759	534	11
XF vs. PA	7750	544	7
XF vs. EC	5213	362	5
XC vs. PA	23300	1563	13
XC vs. EC	13154	908	9
PA vs. EC	23630	1410	7

(The clusters reported only include those for which there are CRPs such that the smaller gene in the pair is at least 20% in size of the larger gene in the pair).

Table 2: Hypothetical genes that are members of CRPs

organism	# of hypo. genes
XF	7
XC	16
PA	19
EC	19
total	61

Of the 61 genes listed above, 12 appeared in more than one pairwise genome comparison.

Table 3: Candidate gene fusion/fission events

organisms	fus in G (fis in H)	fis in G (fus in H)
XF vs. XC	125	257
XF vs. PA	191	277
XF vs. EC	145	220
XC vs. PA	1088	803
XC vs. EC	527	489
PA vs. EC	914	828

Table 4: Candidate run fission (in G)/fusion (in H) events. T is the number of genes between the two series of consecutives genes in G.

organisms	# events	mean T	max T	min T
XF vs. XC	53	611.2	2530	1
XF vs. PA	47	800.3	2343	1
XF vs. EC	21	746.4	2760	1
XC vs. PA	278	1365.3	3870	1
XC vs. EC	57	997.3	4259	1
PA vs. EC	211	1668.6	5557	1

Table 5: Candidate run fusion (in G)/fission (in H) events. T is the number of genes between the two series of consecutives genes in H.

organisms	# events	mean T	max T	min T
XF vs. XC	80	829.8	3692	1
XF vs. PA	69	1637.1	4942	1
XF vs. EC	36	979.4	3571	1
XC vs. PA	505	1724.0	5347	1
XC vs. EC	166	1449.1	4246	1
PA vs. EC	214	1441.9	4140	1

These results are currently being analyzed. In particular we expect to determine more candidate related pairs by inspecting the run fusion/fission events as well as runs that have isolated matches close by.

Previous works related to the results presented here are referenced below.

1. S. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25:3389–3402, 1997.

2. W. Fujibuchi, H. Ogata, H. Matsuda, and M. Kanehisa. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. Nucleic Acids Research, 28:4029–4036, 2000.

3. M. Y. Galperin and E. V. Koonin. Who's your neighbor? New computational approaches for functional genomics. Nature Biotechnology, 18:609–613, 2000.

4. H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Research, 28:4021–4028, 2000.

5. R. Overbeek, M. Fonstein, M. D'Souza, G. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. PNAS, 96:2896–2901, 1999.